View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CORE

Wright State University

CORE Scholar

Browse all Theses and Dissertations

Theses and Dissertations

2009

Type I Error Rates and Power Estimates for Several Item Response Theory Fit Indices

Bradley R. Schlessman
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

 Part of the Industrial and Organizational Psychology Commons

Repository Citation

Schlessman, Bradley R., "Type I Error Rates and Power Estimates for Several Item Response Theory Fit Indices" (2009). *Browse all Theses and Dissertations*. 805.
https://corescholar.libraries.wright.edu/etd_all/805

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

TYPE I ERROR RATES AND POWER ESTIMATES FOR MULTIPLE ITEM RESPONSE THEORY FIT INDICES

A dissertation submitted in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

By

BRADLEY ROBERT SCHLESSMAN
M.S., Wright State University, 2005
B.S. The Ohio State University, 2003

2009
Wright State University

WRIGHT STATE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

December 13, 2009

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Bradley Robert Schlessman ENTITLED Type I Error Rates and Power Estimates for Multiple Item Response Theory Fit Indices BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

_____

David LaHuis, Ph.D.
Dissertation Director

_____

John Flach, Ph.D.
Department Chair

_____

Joseph F. Thomas, Jr., Ph.D
Dean, School of Graduate Studies

Committee on Final Examination

_____

Nathan Bowling, Ph.D.

_____

Corey Miller, Ph.D.

_____

Scott Watamanuik, Ph.D

ABSTRACT

Schlessman, Bradley Robert. Ph.D., Department of Psychology, Wright State University, 2009. Type I Error Rates and Power Estimates for Multiple Item Response Theory Fit Indices.

Despite frequent use of the adjusted chi-square to degrees of freedom ratio (χ^2/df) test for Item Response Theory fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995), there remains a lack of empirical testing of the statistic's Type I error rates and power. The present study compared the adjusted χ^2/df test to two other commonly used IRT fit statistics. The other fit indices examined were S-χ^2 (Orlando & Thissen, 2000) and χ^{2*} (Stone's, 2000). This study also addressed misfit based on the possibility that the item responses analyzed were created based on a different response process than that assumed by the IRT model used to analyze the data. Results suggest that the adjusted χ^2/df test without cross validation has the best Type I error rate, is the test least affected by changes in sample size and test length, and is best suited for the detection of misfit based on violations of the local independence assumption. Stone's χ^{2*} however appeared to be the best statistic to detect misfit based on the model misspecification introduced. Furthermore, the power/Type I error rate trade off for the adjusted chi-square to degrees of freedom ratio test demonstrated that the cut off value for acceptable fit of 3.0 may not always be the ideal cut-off value.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Item Response Theory (IRT) methods continue to be used more and more often for a number of testing applications such as item banking, adaptive testing and test development. IRT methodology provides ability estimates for tests which are generally more informative than scores derived from classical test theory (CTT) methods. CTT methods tend to provide scores which are specific to the set of questions asked in the given test such as the summed score, percentage correct, or percentile scores. These types of scores provide information about an individual's standing on the administered test and only general information about the difficulty of each test item. IRT methods are more informative because item parameters are not sample dependent. The person parameters estimated by the model are not specific to the set of items used on the test, and measurement precision is not assumed to be constant; that is to say that IRT methods allow researchers to calculate conditional standard errors of measurement (Chernyshenko, Stark, Chan, Drasgow & Williams, 2001). This means that unlike CTT methods, in which the standard error of measurement applies to all scores in a particular sample, IRT methods allow the standard error of measurement to differ across scores (or response patterns) while still generalizing across populations. IRT methods also allow for the development of computer adaptive testing, item banking, and

a more informative method of test equating. They also provide information about individual test items and individuals taking the test. IRT's usefulness beyond that of CTT models and the associated differences between the two testing models was outlined in full detail by Embretson and Reise (2000).

This flexibility allows IRT methodology to evaluate if, and more importantly how, items and tests function differently between two groups. This process is often referred to as differential item or test functioning (Ellis, 1989; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Stark, Chernyshenko, Drasgow, 2006). IRT models and analyses have been adjusted to a number of specific types of testing situations due to the large number of test types which can benefit from IRT analyses (Chernyshenko, Stark, Drasgow & Roberts, 2007; Rasch, 1960; Fischer, 1973, Masters & Wright, 1996; Roskam, 1997). Overall, IRT's use as an analytic tool comes from the user's ability to evaluate item and person parameters independently of each other. Researchers can then use this information to make sound decisions on test development and application on both a very large and a very small scale.

For example, if a set of 50 cognitive ability items is administered to a sample of 500 individuals and then that test is scored using CTT methods, a standard form of reporting the results would be to provide the percent correct out of the 50 questions administered. Using CTT methods, also available would be information on which items were answered incorrectly more often than others, and perhaps a significance test value. Beyond those and other such scores derived from those numbers, there would not be much information available on the individuals who took the test or the items within the test itself.

By using an IRT methodology, it is possible to identify how trait level and item properties relate to an individual's item response pattern, and ultimately the probability of that person getting that question correct. Given this information, a latent trait ability estimate which is not linked to the specific set of questions being asked can be calculated. This latent trait ability level, also known as theta (θ), identifies the individual's standing on a latent trait ability continuum which is typically standardized with a mean of zero and a standard deviation of 1.0. Although the range of theta is not restricted when estimating ability, the range of theta values is typically scaled between -3.0 and 3.0 when plotting values because of the rarity of observing individuals whom fall outside of three standard deviations of the mean on any given tested trait or ability. Information about the difficulty and discrimination level of each individual test item is also obtained from IRT analyses allowing for a more informative view of how each item within the test as a whole operates.

The two item parameters most often used in IRT analyses are the item difficulty and item discrimination values. The difficulty of an item is represented by the item's position on a latent trait continuum. Associated with an item's difficulty is the item's point of inflection. The point of inflection is the point on the curve where the rate of change shifts from accelerating increases to decelerating increases. The relationship between an item's difficulty and an individual's trait level can be demonstrated by differing points of inflection for items with different difficulties. Points of inflection which fall higher on the latent trait continuum represent more difficult items. The trait level also indicates the point at which someone is as equally likely to pass an item as they are to fail it. For example, when an individual has a trait level of 1.0, the probability of

that person passing (endorsing an item / providing the correct response) an item with a difficulty of 1.0 would be .50.

The second main item parameter is the discrimination parameter. The discrimination of an item relates to the fact that items within a test may not be equally indicative of a person's standing on the latent trait. Or put simply, some items discriminate between individuals with either high or low ability levels better than other items. The item's discrimination is indicated by its relative slope. These two item parameters are used to estimate each individual's ability level. Figure 1 demonstrates an item characteristic curve for an item with a difficulty of 0, and a discrimination of 5.0. This type of detailed item information allows for a more thorough investigation of the properties of a test and how individuals respond to the questions within it.

The advantages of IRT methods over CTT methods are limited however by certain restrictions IRT methods have which CTT methods do not face. There are stricter assumptions within an IRT methodology than there are for CTT methods. For example, IRT models assume the data are unidimensional and items have local item independence. Unidimensionality occurs when all of the items within a test measure one and only one construct. Insuring this aspect of the test helps to ensure local item independence. Items have local independence if once all of the parameters in the IRT model have been accounted for, there is no further relationship between the items which might influence item responding. These two are closely related and unidimensionality is often seen as an indication of local item independence. Another assumption is that the model accurately represents the data, and that the data follow the form specified by the model. If these assumptions are not met, then any inferences made about the items, tests or ability levels

of the individual's who took the test can be misleading. It is therefore necessary to have a standard means of assessing these assumptions which would allow researchers to reliably evaluate the fit between the model they are using to analyze their data and the data itself.

More attention needs to be paid to the fundamental issue of model fit in order for the benefits of using IRT to become fully realized in applied settings. Model fit simply means that the mathematical model being used to analyze the data accurately represents the data being analyzed. For example, in structural equation modeling (SEM) there are a number of standard fit indices with known cut-off values (Hu & Bentler, 1998). These cut-off values set a minimally necessary level of model fit to the data which need to be met before it can be assumed that the model accurately represents the data. In SEM, fit is generally evaluated by a number of different indices. This means that three to four fit indices are computed (e.g. Chi square test, Root Mean Square Error of Approximation, Comparative Fit Index, & Standardized Root Mean Residual) and acceptable fit is generally evaluated by comparing the values associated with the current model to preset cut-off standards. If a majority, or three out of four, of the fit indexes have acceptable levels of fit as determined by the agreed upon fit statistic cut-off values, it is assumed that the data itself is represented well by the proposed model (Kline, 2005).

The evaluation of fit for IRT models has not experienced the same level of standardization and consensus on methodology as SEM. There has been an overall lack of consensus on what fit statistic or technique best evaluates model fit for IRT models. A number of statistical procedures have been developed and utilized to evaluate item fit for IRT models. Some of the first attempts to evaluate the fit of IRT models were general chi-square fit indices such as Yen's Q_1, and Bock's χ^2 (Bock, 1972; Yen, 1981).

Limitations of these models led Orlando & Thissen (2000) to create a form of the Person χ^2 index, S – χ^2. Stone (2003) suggested that traditional χ^2 goodness of fit evaluations of IRT models may not be appropriate for shorter tests due to imprecise estimation of the underlying latent trait. Stone created the χ^2* statistic in order to account for this issue. Another commonly used IRT fit statistic in organizational settings is the adjusted χ^2 to degrees of freedom (adj. χ^2/df) ratio test (Drasgow, Levine, Tsien, Williams and Mead, 1995). This fit statistic has been used in numerous organizational studies due to its simplicity and ease of use (LaHuis & Copeland, 2009; Scherbaum, Cohen-Charash & Kern, 2006; Stark, Chernyshenko, & Drasgow, 2005; Stark, Chernyshenko, Drasgow & Williams, 2006). This method was created in order to address the sensitivity to large sample sizes standard χ^2 fit statistic tests demonstrate. Each of these techniques and or models for assessing model fit possesses both benefits and limitations. In general researchers and individuals in applied settings tend to use the index of fit which is most familiar to them, or which is most convenient. Often the adj. χ^2/df test is used due to a lack of knowledge about the other indices, or how to actually calculate them.

The present study investigated the tradeoff between Type I error rates and power among three commonly used IRT fit statistics for dichotomously scored items. The three fit statistics investigated were, Orlando and Thissen's S- χ^2, Stone's χ^2*, and the adj. χ^2/df ratio test. These fit indices were evaluated under a number of conditions which may influence the detection of model misfit. In the following sections I will review briefly the history of IRT, identify the major differences between IRT and CTT, identify the assumptions associated with IRT analyses, review three often used IRT models, and identify the IRT fit indices evaluated.

History of IRT

Thurstone (1925) wrote *A Method of Scaling Psychological and Educational Tests* which contained many of the same concepts used in modern IRT methods. He plotted points corresponding to age and the proportion of correct responses on a children's ability test. These plots roughly fit the pattern of the S-shaped curves associated with IRT analyses. Thurstone's model and IRT models both express the probability of success on an item as the function of a variable associated with the individual. His analysis also represented both the item locations and the individual responses on the same scale as the variable of interest allowing for simultaneous evaluation of each. Richardson (1936) continued this sort of work when he derived the relationship between IRT model parameters and classical item parameters. In 1943, Lawley published a paper which showed that many of the constructs of CTT could be expressed in terms of the parameters associated with item characteristic curves. These are curves which relate item responses to the probability of correctly answering an item on some latent trait continuum. Another individual influential in IRT was Tucker (1946), who furthered the development on linking classical test theory parameters with the parameters used in item response theory.

One line of research that led to consistent interest was that of Lord and Novick (Lord, 1953; Lord & Novick, 1968). Lord and Novick's textbook, *Statistical Theories of Mental Test Scores* (1968) along with a paper by Lord (1953) were driving forces behind the development of theory and application for item response theory. Lord's work applying the normal-ogive model to real test data was perhaps the most influential breakthrough in the field, and led to a number of technical reports by Birnbaum, (1957,

1958a, 1958b) whom replaced the normal-ogive curves with logistic curves and also developed the methodology and equations needed for other statisticians to implement these models.

A separate line of IRT research was developed in Denmark by Rasch (1960). He was particularly interested in creating a model in which person and item parameters were completely separate. He called this aspect of his models *specific objectivity*. Rasch inspired a number of students and colleagues, and they subsequently utilized and improved upon his model. Dr. B.D. Wright of the University of Chicago continued work on the Rasch model and was instrumental in bringing knowledge of this methodology to practitioners.

From these models, IRT methods have been expanded to encompass a number of methodological and conceptual domains. Although these techniques existed, many individuals were not able to utilize them due to a number of conditions which these methods require. The models themselves are computationally demanding to run and until the somewhat recent availability of cheaper, more powerful computers, utilizing IRT methods was not a practical option for everyday researchers or practitioners. Even with the availability of computer software and hardware to run these models, they still require large amounts of data in order to obtain stable findings, and there are a number of requisite assumptions necessary for IRT models in order to obtain reliable and valid results. A more complete history of developments in mathematics and theory leading up to current day IRT methods can be found in Bock (1997).

Advances in the field have led to a proliferation of varying models which utilize the IRT framework. Proof of the increase in interest can be seen by special issues of

Journal of Educational Measurement (1977) and *Applied Psychological Measurement* (1982) devoted to IRT, and entire books by Lord (1980), and Hambleton and Swaminathan (1985). Since that time, computational power has reached the demands of the models and IRT has enjoyed an expansion into new fields and an increase in use by general practitioners. More recently, Embretson and Reise (2000) wrote a book especially for psychologists.

Many models have been developed for a vast array of situations. A partial list of these models includes two and three parameter unidimensional and multidimensional logistic models for dichotomous and polytomous data (Reckase, 1997; Rost & Cartstensen, 2002; McDonald, 2000), a Graded Response Model for Likert-type personality data (Samejima, 1969), a model for multidimensional longitudinal data (te Marvelde, Glas, Van Landeghem & Van Damme, 2006), unidimensional and multidimensional ideal point models (Maydeu-Olivares, Hernandez, & McDonald, 2006; Chernyshenko, et al., 2007), and unfolding IRT models (Roberts & Laughlin, 1996). Despite the large number of available programs and applications of IRT models, research in the area of model fit for IRT models lags behind development and usage of these models.

Differences Between IRT and CTT

CTT has been employed for psychological measurement in one form or another since the beginning of the 20th century. Spearman's work on test theory, reliability and validity (Spearman, 1904, 1910) laid the groundwork for further developments in the area of test development. CTT began as the standard for test development in the 1930's when standardized testing became popularized, and is still used today for many psychological

tests. It is widely recognized that Gulliksen's book, *Theory of Mental Tests* (1950), was the defining book of its time on test theory, but even he stated "Nearly all the basic formulas that are particularly useful in test theory are found in Spearman's early papers."

However, in recent years CTT has been supplemented in some areas by IRT and IRT is often used by many large scale testing organizations as the newer and more useful way to develop tests and assessments. Lord and Novick's book (1968), which introduced model based measurement, was the impetus for this change. Large scale change has been slow and many schools still teach CTT as the main theory for test development and scoring. IRT, which is a latent trait theory, creates a score which is based on an individual's responses and the properties of each individual item. IRT was at first attractive to test developers as a means to investigate item bias. Of particular interest was the potential bias cognitive ability tests demonstrated towards certain minority groups. It was suggested that perhaps the tests contained questions which might operate differently for the minority groups than they did for the white majority due to their content. The ability to identify specific items which may be interpreted or understood differently for minority test takers than it is for a white test taker would allow test developers to remove or alter those items which were unfairly biased against minority test takers.

IRT methods follow a very different set of rules for measurement of ability than CTT. While CTT is basically an arithmetic sum of a person's recorded score plus the error associated with that score, IRT is a model-based approach which accounts for both the individual and the scale and their corresponding parameters which can differ from item to item, test to test, and person to person. These newer methods emerged as a more informative way of testing which utilizes a completely different theoretical basis for

measurement than the CTT methods which were used beforehand. Embretson and Reise (2000) listed in full detail the ten main differences between CTT and IRT methods. The following is a brief review of the more general differences in rules of measurement between IRT and CTT as described by Embretson and Reise. The complete list of ten differences is provided in Appendix A. See Embretson and Reise (2000) for an extended explanation of these differences and the complete list of rule differences.

The first difference deals with the standard error of measurement. Under CTT there is an assumed consistency in regards to the standard error of measurement within a population. Under IRT modeling standard errors differ across scores within a population but generalize across populations. This allows for a more exact measurement of the standard error of measurement which is important for both describing the psychometric quality of a test and individual score interpretations. The second difference between CTT and IRT models stated by Embretson and Reise is the idea that longer tests are more reliable. This idea was based on measures of reliability such as alpha, such that if items were highly correlated, they would be considered more reliable. According to CTT by including a large number of questions and ensuring consistent answers on multiple questions asking about the same concept or trait, the consistency of answers could be obtained. In contrast, IRT purports that by identifying latent trait levels and measuring those instead of raw scores via adaptive testing methods, the reliability of scores can be ensured with shorter tests.

Another difference between the two testing methodologies is that it is assumed that unbiased estimates of item properties may be obtained from non-representative samples in an IRT framework, while CTT methods cannot do this. This is because in

CTT item difficulty is measured by a p-value, and discrimination under CTT is the item total discrimination based on an entire population's responses. These can differ substantially if measured with unrepresentative samples. For example, if a group of 2,000 individuals are split at the median of performance on a task and then have their performance graphed, a linear representation of their performance can be created. It is possible however that a greater difference exists among individuals in the higher performing group than among individuals in the lower performing group. IRT methodologies are able to more accurately capture such differences.

Finally, when using classic test theory, test scores are compared to a normalized distribution of scores. If 1,000 individuals take a test, a normal distribution of scores is expected to be obtained. Under CTT comparing one individual score to the entire population is how a person's standing on the test is evaluated. IRT test scores do not obtain meaning from comparing test scores to a norm, but from comparing latent trait levels of individuals to specific item's difficulty and discrimination values. These represent some of the main differences between CTT and IRT methodology which allow IRT methods to be used in a more effective way to gather and use information.

Assumptions of IRT Models

Mathematical models make assumptions about the data which need to be met in order to be able to correctly estimate the parameters within the model. When the assumptions hold, the model is able to correctly estimate the relationship between the observed and unobserved variables or constructs in the model. CTT methodology makes only a few weak assumptions about the data used to estimate the parameters of the model. CTT's basic form is:

$$TTS = O + E, \quad (1)$$

where TTS is the True Total Score, O is the observed score and E is the error associated with the measurement of O. CTT has three basic assumptions. The first is that the average error across subjects is zero. The second assumption is that error is not related to other variables. Finally, errors are assumed to be normally distributed and homogeneously distributed across individuals.

Alternatively, IRT models make a number of strong assumptions about the data. The two main assumptions of item response modeling as stated by Embretson and Reise (2000) are that of unidimensionality and that item response functions (IRF) have a specific form. A third assumption not explicitly stated as a major assumption by Embretson and Reise is the assumption of local independence. A final assumption is that the distribution of ability within a population is standard normal.

The first assumption made by IRT models is that of unidimensionality. This is the concept that only one latent trait is being measured at a time. The latent trait is an unobserved characteristic which is presumed to be responsible for observed scores. This is represented by "θ" (theta). For estimation purposes, theta is often scaled with a mean of zero and a standard deviation of one.

Unidimensionality of the latent trait is usually not a large problem because performance on any one particular aspect of a test can be assumed to be accounted for by a single latent construct which usually obtains at least minimally acceptable levels of unidimensionality. As explained by Hambleton and Swaminathan (1985), if a test is not unidimensional, it may function differently within separate subpopulations (i.e. different cultures) and not provide the same results at a given ability level. This would be due to

the effect that the other variables or abilities being inadvertently measured are having on the subjects' scores. The achievement of local independence means that once all of the parameters in the IRT model have been accounted for; there is no further relationship between the items. It has been shown IRT analyses are robust to relatively small violations of local independence (Glas & Hendrawan, 2005).

The second main assumption of IRT models is that the IRF which represents the data has a pre-specified form. The curve relates changes in trait levels to changes in the probability of a specified response. The IRF is a nonlinear regression of the probability of success on trait level. This creates an "S" shaped curve which monotonically increases as a function of the individual's trait level, as displayed in Figure 1. As can be seen in Figure 1, in the middle of the curve small changes in theta imply large changes in item solving probability, or providing a correct response, and at the extremes of the curve, large changes in trait level lead to small changes in the probability of a correct response. The specified shape of the curve is determined by a function which relates the person and item parameters to the probabilities of responding. Not all IRT models produce S-shaped curves. For example, unfolding model's curves have more of an inverted u shape.

The third assumption refers to the fact that responses to each item need to be independent of all other questions within a test. This assumption states that the relationships between the items and individuals completing the assessment is completely characterized by the IRT model and are statistically independent. This relationship is contingent on item responses being conditioned on an individual's trait level and item difficulties. It has been noted that unidimensionality and local independence are closely related. The achievement of local independence can be evidence for unidimensionality if

the model contains person parameters on just one dimension. Although closely related, unidimensionality and local independence are two different concepts. The final assumption simply states that the distribution of ability within a population is standard normal, which leads to the scaling of the ability parameter within IRT to also be standard normal.

Ability Estimation

As stated above, the estimation of ability is fundamentally different in IRT models compared to that of CTT models. A person's trait level is not a simple sum of item responses, but is instead an estimation of an individual's latent trait level. By examining a known response pattern, the question asked is essentially "what trait level is most likely to explain this particular set of responses." Given a test with very difficult questions, what is the likelihood that a person with a low, medium or high ability level would get most of the questions correct? Individuals with higher ability levels would be expected to score higher on tests with more difficult questions. IRT methods require a search process for the ability level that yields the highest likelihood for the observed responses.

A common method of estimating ability is the maximum a posteriori (MAP) scoring. This method uses a Bayesian estimator to derive an ability estimate. The first step is to identify a set of density weights for a given number of theta values. For example, it is common to calculate density weights for values ranging from -3 to 3 in .10 increments. The density weights are typically based on a standard normal distribution. This serves as the prior distribution. A posterior distribution is calculated by multiplying the prior distribution by the likelihood of the item response pattern given a theta level.

That is, the posterior distribution represents the joint probability of theta and the likelihood of the response pattern given theta. MAP estimates are the ones that maximize the posterior distribution.

IRT Models

One Parameter Logistic Model

IRT is a methodology based on the concept that individual items are only indicators of a higher order latent factor which is what is of interest when testing, and is what drives item responses. The Rasch model, also known as the 1 Parameter Logistic Model (1PLM) as it only measures the difficulty level of the test item, predicts the probability of person *j* on item *i* answering an item correctly as:

$$P_{ij}(Y=1|\,\theta_j,\,\beta_i) = \frac{\exp[(\theta_j - \beta_i)]}{1+\exp[(\theta_j - \beta_i)]}. \qquad (2)$$

Embretson and Reise (2000) pointed out three main features of the Rasch model which are displayed in Figure 2. Figure 2 shows three items with the same discrimination values, but with three different difficulty levels and is helpful in understanding the three main features of Rasch models. First, the probability of successfully answering a question gradually increases as trait levels increase for each item. Second, as can be seen in Figure 2, when the items differ only in difficulty, and the slopes of the curves are equal the curves do not cross. Finally, the point of inflection of the IRF, the point at which the rate of change shifts from accelerating increases to decelerating increases, takes place when the probability of correctly answering an item is .50. At this point, a person is as likely to answer an item correctly as they are to answer it incorrectly. This point is labeled P in Figure 2.

Two Parameter Logistic Model

The two-parameter logistic model (2PLM) is a simple enhancement of the one parameter model. By adding a discrimination parameter, α, the probability that a person *j* solves item *i* is given as:

$$P_{ij}(Y=1|\,\theta_j,\,\beta_i,\alpha_i) = \frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[\alpha_i(\theta_j-\beta_i)]} \quad (3)$$

It is important to note that equation 3 is equivalent to equation 2 under the circumstance that α is equal to 1.0 for every *i*. This would have the effect of making α a constant and not allowing the discrimination parameter to freely vary at the item level. The 2PLM is a more realistic model for most tests as it is difficult to show that each item on a test demonstrates the same level of discrimination. As displayed in Figure 3, once the discrimination parameter has been freed, the slopes of the individual IRFs can differ substantially, and it can be inferred that a constant item discrimination would not fit the data as well as allowing that parameter to vary. Figure 2 displays three items which all have discriminations of 3 but differ in difficulty. The items have difficulties of -1.0, 0.0, and 1.0 for items 1, 2, and 3 respectively. The items represented in Figure 3 allow the discrimination parameter to vary. Item 1 has a discrimination of 1.0. Item 2 has a discrimination of 5.0 and item 3 has a discrimination of .75. As can be seen in Figure 3, the larger the discrimination, the steeper the IRF, which means that small increases in ability translate into large increases in the probability of a correct answer. On the other hand, for items with smaller discrimination values, the IRF will be less steep showing that small increases in ability will lead to small increases in the probability of a correct answer for this item.

Three Parameter Logistic Model

A three parameter logistic model (3PLM) is very similar to a 2PLM except for the addition of the quasi guessing parameter (γ). The guessing parameter sets a lower limit for the IRF at some point above zero. The parameter is based on the concept that for any multiple choice question, an individual with zero ability has a given probability of getting the question correct. For example a multiple choice question which has five options should have a lower asymptote around .20 as the individual answering the question has a 1 in 5 chance of randomly guessing the correct answer. The 3PLM can be represented by:

$$P_{ij}(Y=1|\,\theta_j, \beta_i, \alpha_i, \gamma_i) = \gamma_i - (1-\gamma_i)\,\frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[\alpha_i(\theta_j-\beta_i)]}, \qquad (4)$$

where α_i and β_i still represent the discrimination and difficulty parameters for item *i* respectively and γ_i represents the lower asymptote or guessing parameter of item *i*. However, estimates of lower asymptotes for a 3PLM often differ from the random guessing probability. This is because for some questions examinees may be able to eliminate certain distracters without knowing the correct answer, thusly increasing their probability of correctly answering the question. Estimation problems for item parameters have been known to occur if questions within a test have different lower asymptotes, so it is common to estimate a common lower asymptote or constrain all the lower asymptotes to all be equal.

The previous three models are examples of dominance IRT models. Dominance models are not however the only type of IRT model. Models which utilize an ideal point methodology are not currently used as often but are becoming increasingly more popular, especially for use on attitude or personality type data. Ideal point models are based on the concept that non-endorsement of an item is not necessarily a function of not having a high enough level of a certain trait. The ideal point methodology is instead based on the

concept that it is possible that individuals do not endorse an item because it either represents too much or too little of the given construct. In this study, an ideal point model was used to generate data that violate assumptions made about IRFs and may cause model-data misfit. This model will be explained in more depth in the methods section.

IRT fit indices

As stated earlier, an IRT model can only be demonstrated to be useful if it can be shown that the model fits the data being analyzed. Yen (1981) created a standard χ^2 statistic as a way to demonstrate the necessity of model-data fit. Yen's procedure consists of placing individuals into 10 separate groups based on their theta values and within each group, calculating the proportion of the group which answered correctly and the proportion *predicted* to get the question correct. The differences between those two proportions are squared and then multiplied by the cell size and then summed in order to form the χ^2 index. This can be represented by the formula

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \quad (5)$$

where,
N_j = the number of examinees in cell *j*, and
O_{ij} = the observed proportion of examinees in cell *j* that endorses item *i*.
The term E_{ij} is the predicted number of examinees in cell *j* that endorses item *i* and is computed by

$$E_{ij} = \frac{1}{N_j} \sum_{k \in j}^{N_j} \hat{P}_i(\hat{\theta}_k). \quad (6)$$

In equation 6, $\hat{P}_i(\hat{\theta}_k)$ is the item characteristic function for item i, which is evaluated using the trait estimate, θ_k for examinee k, and the item parameters estimated for item i.

The summation of which is taken over examinees located in cell j. $\hat{P}_i(\hat{\theta}_k)$ can be represented as:

$$\hat{P}_i(\hat{\theta}_k) = \hat{c}_i + \frac{1 - \hat{c}_i}{1 + e^{-1.7\,\hat{a}_i(\hat{\theta}_k - \hat{b}_i)}}. \quad (7)$$

The terms $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ represent the estimated discrimination, difficulty and guessing parameters respectively for item i. For a two parameter model $\hat{c}_i$ is fixed to zero, and for a one parameter model $\hat{c}_i$ is fixed to zero and $\hat{a}_i$ is fixed at a constant for all items.

Bock (1972) proposed a similar procedure for the nominal response model except that his model provided estimates for item parameters for all of the answer choices given with the item, not just the correct choice. This model can be restricted to deal only with the characteristic function of the correct answer. When Bock's model is restricted in this way it is equivalent to Yen's except for two differences. First, examinees are grouped into J cells, but J does not have to be equal to ten. Secondly, in the calculation of the expected scores Bock uses a median theta value for examinees in cell J. The differences this model and the one proposed by Yen are generally trivial.

Some of the problems which have been identified with the chi-square tests mentioned are that their validity is questionable when expected values are less than one, they are sensitive to sample size and test length, and the creation of the theta or ability subdivisions (strata) is usually done in an arbitrary way (Hambleton & Swaminathan, 1985; Reise, 1990; Stone, 2003). For example, Orlando and Thissen (2000) found that Q_1 exhibited an inflated empirical Type I error rate as high as 0.96 for a given nominal rejection rate of $\alpha = 0.05$ on a test with ten dichotomous items. Muraki (1997) commented that if the number of intervals is too large, the chi-square statistic can become

inflated. More recently Stone (2000) found that the number of ability subgroups does affect the performance of the fit statistic. It has also been suggested that although χ^2 fit statistics are probably the most widely used of all of the model-fit assessments, conclusions based solely on them are often avoided due to their sensitivity to sample size and insensitivity to certain forms of model-data misfit (Chernyshenko, et al., 2001). For these reasons Yen's and or Bock's χ^2 will not be simulated in this study as they are known to have poor psychometric qualities. The three fit statistics mentioned below will be simulated in this study. All three fit statistics follow the same general formula, what differentiates these models from traditional χ^2 statistics and between themselves are the different ways in which the observed and expected values are generated.

Orlando & Thissen's S - χ^2

Orlando and Thissen (2000) proposed a χ^2 index with the form

$$\text{S - } \chi^2_i = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1-E_{ik})} \ , \tag{8}$$

where the observed proportions (O_{ik}) for item i and the number correct score for group k are computed from the data. The expected proportions (E_{ik}) are computed using the joint likelihood for the number correct score k for all of the items, where S_k is the number correct score posterior distribution for score group k, T_{last} is the IRF for the last item, S^*_{k-1} is the number correct score posterior distribution for score group k -1 without the last item, and S^*_k is the number correct score posterior distribution for score group k without the last item. This is represented in equation 9.

$$S_k = T_{last} S^*_{k-1} + (1- T_{last})\ S^*_k \ , \tag{9}$$

This recursive algorithm is repeated for all items, omitting a different item for each iteration. This provides the joint likelihoods for each score group without item *i* (S_k^{*i}). Using the number correct score likelihoods they are combined with each of the omitted items. This provides the desired proportions of examinees with score *k* who responded correctly to item *i*:

$$E_{ik} = \frac{\int T_i S_{k-1}^{*i} \phi(\theta) d\theta}{\int S_k \phi(\theta) d\theta}. \quad (10)$$

Orlando and Thissen approximated the integrals used in Equation 10 using rectangular quadrature over equally spaced increments of θ from -4.5 to 4.5.

One of the main advantages of *S* - χ^2 over Q_1 is that the *S* - χ^2 procedure is based on actual test scores (the number correct), whereas Q_1's grouping procedure relies on sample and model dependent cut scores. Orlando and Thissen (2003) further investigated the *S* - χ^2 index and found adequate Type I error rates at test lengths of 10, 40, and 80 for dichotomous items and a sample size of 1,000. Type I error rates were also estimated for one, two, and three parameter logistic models. They found that the *S* - χ^2 index displayed empirical Type I error rates between .04 and .07 at an α level of .05. Finally, Orlando and Thissen (2003) also found that the empirical power of *S* - χ^2 improved as sample size increased from five hundred to two thousand.

*Stone's χ^2**

Agresti (1990) observed that when expected frequencies are a function of *t* parameters, the nominal *df* are decreased by *t.* Yen (1981) originally suggested that although expected values for item responses are dependent on item parameters, ability estimation is based on the interaction all of the items. Accordingly, given a long test and

the fit for any one particular item, the loss of *df* would be negligible due to the estimation of ability. Stone (2000) commented that the method of considering the loss of *df* negligible essentially treats ability as a known. Stone considered this a problem especially in the case of shorter tests in which the precision of the estimation of ability is often suspect, which can lead to classification errors. Stone suggests that rather than use point estimates of ability, it is possible to consider the uncertainty associated with the estimation of theta for tests with a small number of items. He substituted multiple expectations conditional on a model which relates the unknown quantity to data that is observed for the unknown quantity (theta). For the estimation of his χ^2* statistic, Stone approximated the continuous theta scale by a set of discrete points, and probabilities of the response at each score level given the theta levels by

$$r_{jk} = \sum^{N} x_{jn} P(x_n | X_k) A(X_k) / P(x_n), \qquad (11)$$

where, r_{jk} is the posterior expectation of an item for score response level *j* and theta level *k*; *n* refers to the number of examinees in the sample; x_{jn} is equal to 1.0 if the observed response of the *n*th examinee for the item equals *j* and is 0 otherwise; $P(x_n | X_k) A(X_k)$ is equal to the conditional probability of the *i*th examinee's response pattern to (x_n), given the ability level X_k. Finally, $P(x_n)$ is the marginal probability of observing response pattern x_n for an examinee with an unknown theta value from a population in which theta is normally distributed.

Stone (2003) measured Type I error rates and empirical power for the χ^2* fit statistic and found acceptable Type I error rates for tests with lengths of 6 and 12 given a normal ability population. Acceptable Type I error rates are below .05. In regards to

empirical power, as expected, power decreased as α went from .10 to .01 and increased as *n* increased from 500 to 2,000. He also manipulated the ability distributions and found that when a skewed ability distribution was used, Type I error rates were found to significantly increase.

Adjusted Chi Square / Degrees of Freedom Ratio Test

Although a number of the aforementioned techniques have had extensive testing and evaluation, an often used test of model fit which has not been investigated in-depth is the adjusted χ^2/df ratio test. This test is commonly used for IRT fit analyses in organizational research (Bolt, Hare, Vitale, & Newman, 2004; Stark et al. 2006; Zickar, Russell, Smith, Bohle, & Tilley, 2002). Drasgow et al. (1995) proposed a χ^2statistic which is adjusted in order to account for the sensitivity to sample size and to allow comparisons between different samples, tests and parameter differences. By adjusting the χ^2 statistic to the magnitude that would be expected in a sample of 3,000, the large sample sizes necessary to run these analyses do not influence the χ^2 to degrees of freedom statistic as much. Despite the popularity of the adjusted χ^2/df ratio test (Bolt, et al., 2004; Stark et al. 2006; Zickar, et al., 2002), there is little empirical research concerning it. In the present study, I examined how a number of conditions affect the Type I error rates and power estimates of this widely used fit statistic.

The formula for the adjusted χ^2/df ratio test is

$$\chi_i^{\,2} = \sum_{k=1}^{s} \frac{[O_i(k) - E_i(k)]^2}{E_i(k)}, \qquad (12)$$

where *s* represents the number of keyed options, $O_i(k)$ is the observed frequency of endorsement of item k, and $E_i(k)$ is the expected frequency of endorsement of item k. Each χ^2 statistic is then adjusted to the magnitude that would be expected in a sample of

3,000 and divided by the number of degrees of freedom. According to Drasgow et al. (1995) a ratio of more than 3.0 is viewed as a sign of model-data misfit.

The expected response distribution is based on item parameter estimates and ability levels which represent the discrete ability subgroups (Drasgow, et al. 1995). The expected response distribution is calculated using

$$E_1(k) = N \int P(v_i = k \mid \theta = t) f(t) dt, \quad (13)$$

in which $f(t)$ is the theta density which is generally taken to be standard normal.

Item Doubles and Triples. χ^2 statistics for single items tend to be insensitive to unidimensionality violations and certain types of misfit. This can be demonstrated by a situation in which an empirical item response function (IRF) is above the estimated IRF at low trait levels and is below it at high trait levels (Figure 4). In this situation a χ^2 test for an individual item will be close to zero. This is because the χ^2fit statistic is a marginal statistic and the estimated IRF is integrated with an entire normal theta density. This problem can be adjusted for by computing the χ^2statistic for pairs and triples of items. For example a twenty item test might have 6 questions from the lower end of the difficulty scale and 7 items from both the middle and high end of the difficulty scale, and a triple would have one item from each range of values. Van den Wollenberg (1982) and Glas (1988) showed that when using a Rasch model, pairs of items instead of singles are more sensitive to violations of local independence. The technique used in this study for calculating the adjusted χ^2statistic for item pairs and triples is the same as the technique used by Stark, Chernyshenko, and Drasgow (2005) in which, they found pairs and triples of personality items fit better than singles.

The expected frequency for a pair of items can be computed as:

$$E_{i,i'}(k,k') = N \int P(v_i = k \mid \theta = t) P(v_i = k' \mid \theta = t)\, f(t)dt. \quad (15)$$

This equation represents a pair of items in the $(k,k')^{th}$ cell of a two-way table for items *i* and *i'*. The observed frequencies are counted in each cell. Item triples are computed in a similar way.

It has been consistently suggested that the best fitting IRT models have adjusted χ^2 to degrees of freedom ratios for item singles below 3.0. Item doubles and triples with ratios below 3.0 are also recommended (Chernyshenko, et al., 2001; Drasgow et al., 1995; Scherbaum, Cohen-Charash & Kern, 2006). The current view in the literature is that if the ratio exceeds 3.0, for item singles, doubles and triples, it can be inferred that the model does not fit the data due to the parametric form of the item/option response function being violated (Stark et al. 2006). There are a number of other reasons which might cause the data to have poor fit, some of which are associated with the underlying assumptions of IRT methodology and the parameters associated with the estimation of the model.

Present Study

Type I Errors

One of the major goals of the present study was to compare the three alternative fit indices in terms of their Type I error rates. Previous research has indicated that under certain conditions, S-χ^2 and Stone's χ^2* have acceptable Type I error rates. It is less clear as to how the adj. χ^2/df ratio test performs. Thus, I sought to evaluate which test exhibits the best Type I error rates. The standard Type I error rate of .05 is generally acknowledged as the desired or 'best' Type I error rate for a fit statistic. It is assumed that there is a trade-off between Type I error and power, such that as one increases, the other

decreases. By allowing a Type I error rate of .05, it is assumed that the fit index will be better able to detect actual differences when they *are* present.

Research Question 1. Which of the three alternative item fit indices exhibits the best Type I error rate?

Causes of Misfit

There are a number of reasons a model may display a low level of model fit. Some of these causes include the number of examinees, differences in the number of items on the test, violations of local independence, violation of the IRF assumption, and abnormal or restricted theta distributions. Each one of these can have an effect on an IRT's analysis and corresponding Type I error rates and power estimates.

Number of Items / Sample Size.

The number of items within a test has a direct effect on the distribution of available thetas, or ability level estimates, for an IRT analysis. For example, if three individuals complete a test which is three questions long, a max number of three response patterns and corresponding theta values, out of eight possible, will be collected. As such, ability estimates are based on all the items within a test, and when the number of items increases, the number of possible theta values for any given individual also increases. A second issue when dealing with test length was addressed by Stone and Hansen (2000) concerning classification errors. It is sometimes desirable to split groups of individuals into ability subgroups or intervals for analysis purposes, such as when graphical fit analyses are run. Stone and Hansen defined classification error as a situation when an examinee who should be assigned to a particular subgroup is wrongly assigned to another subgroup. They commented that this problem is more likely to occur for shorter tests or

assessments because ability estimates are more likely to be imprecise when associated with shorter tests. Imprecision of any kind in ability estimation can lead to incorrect decisions and Type I errors. Given that it is desirable for an item fit index to not be affected by sample sizes or test length, an important question is, "which of the indices is least affected by changes in sample size and test length?"

Research Question 2: Which item fit index is least affected by sample size?

Research Question 3: Which item fit index is least affected by test length?

Local Independence.

One of the ways local independence can be violated is when the answer to one question provides information about another question. This would mean that the probability of answering the second question correctly would be influenced by the information provided in the first question. If at a fixed theta level, scores were not statistically independent, it would be evidence that there was an unmeasured second trait or outside factor giving some examinees higher scores than the others whom have the same ability level. It is assumed that there is no relationship between the items beyond that stated by the model parameters. This does not mean that the items will be uncorrelated. Positive correlations will occur when there are variations in theta levels measured by the test items. Item scores however should be uncorrelated at any given fixed ability level.

Violations of local independence can have an effect on an IRT analyses' ability to estimate the parameters necessary to assess model fit. Yen (1984) found that when sets of items demonstrated violations of the local independence assumption, item parameter estimates, both the discrimination and difficulty parameters, were larger than when

estimates were obtained from sets of items without these violations. By manipulating the number of items within the test which display violations of local independence, its' effect on Type I error rates and power estimates can be assessed.

The goal of IRT analyses is to represent a pattern of responding using a mathematical model. Traditionally, the fit of the model is assessed one item at a time. By doing this, it is assumed that if each individual item fits well, the entire set of items will have satisfactory fit due to local independence. By using item doubles and triples and calculating fit statistics, an explicit test of how successful a model is at predicting patterns of responding is possible.

Research Question 4: Which item fit index best detects violations of local item independence?

Range Restriction of Theta

As noted above, ability level estimates are based on the entire set of items administered. The fit of the data to the model is in turn evaluated based on the entire set of theta values estimated within the sample. The theta values estimated are often based on a standard normal distribution. This produces theta values generally ranging from -3.0 to 3.0 with a mean of zero and a standard deviation of 1.0. If the population from which the theta values are estimated is not standard normal, but instead is restricted in some way, the corresponding estimated theta values would not themselves have a standard normal distribution. This restriction may have an effect on the evaluation of model fit indices' Type I error rates and power estimates.

Research Question 5. Which item fit index performs best when there is a restriction of range in theta?

Model Specification.

Model misfit can also originate from model misspecification. It is possible that when an individual answers a question, he or she is not responding to the question based on the same response model under which the test item was created, and this can lead to a violation of the assumption that item responding follows a pre-specified form. The form which any individual IRF takes shows how changes in trait levels relate to changes in the probability of a specified response. Most tests are created under a dominance response model process in which higher levels of responding on a trait represent an individual having a higher level of that latent trait. As stated by Stark, et al., (2006) "misspecification of the response process can adversely affect the accuracy of personality questionnaire scores and the predictions made concerning the behavior of respondents" (p. 25). The majority of personality scales are developed with a dominance response method in mind and are derived from Likert's (1932) work on scale development. The appropriateness of this assumption was questioned by Stark et al. (2006). They concluded that the use of an ideal point methodology over that of a dominance model methodology provided as good or better fit to personality data due to its' flexibility and ability to account for an ideal point response process.

Differences in the assumptions regarding item responding underlie the differences between the two methods. In a dominance response process if both items and individuals are represented on a continuum, it is assumed that a person will have a very high probability of endorsing a positively worded item when their standing on the latent trait dimension is at, or greater than, that of the item being asked. For example, consider an openness to new experiences item, "I enjoy trying new things." An individual with a very

high level of openness would be assumed to have a very high probability of endorsing that item. The probability of a positive response (endorsing the item) increases as the person's theta level approaches and surpasses the item's difficulty level. This creates a monotonically increasing s-shaped curve as seen in Figure 1.

One alternative to this method of test creation is the use of an ideal point for responding. This response process uses the idea that a person will only endorse an item if it is located near their trait level on the latent continuum. This implies that a person may not endorse an item because it represents either too much or too little of the trait of interest. According to Thurstone's (1928) method, an individual may not endorse an item because they may feel as though their trait level is either too far below the level of the item or too far above it. The distance between where the item stands on the trait continuum and where the individual feels they stand on the same item, in an absolute sense, is what is of most importance. As this distance increases, the probability of item endorsement decreases accordingly. This can lead to non-monotonic bell-shaped IRFs as shown in Figure 5, which demonstrates the different responding patterns between applicants and incumbents on a personality assessment. The ideal point response process implies that individuals who have a moderate level of openness to experience will endorse the item "I somewhat enjoy trying new things" with a higher probability than would individuals at either extreme end of the latent trait continuum. This is due to their location on that continuum in comparison to individuals that are either very high or very low on that latent trait, as the question itself is more of a moderate question. It is possible to envision an individual reasoning that they are either too high on that construct to endorse it, or too low.

The dominance response process works exceedingly well in a number of contexts such as cognitive ability testing (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). This makes logical sense because if an individual is able to answer a simple math problem such as 5 x 5 = X, they will be more likely to be able to provide the correct answer for the square root of 625 than an individual that is unable to correctly answer the first question. For some non-cognitive constructs it is difficult to say with a high level of confidence that an individual will endorse one item based on their response to prior items. This is because non-cognitive items are often prone to a certain level of interpretation, opinion, or preference and are in general non-objective. These individual interpretations may lead to non-endorsement from both above and or below necessitating an IRT model for the evaluation of the responses which can account for this type of responding.

Unlike cognitive ability questions, personality questions have no 'true' right or wrong answer. Often individuals taking these tests are asked to provide the answer that best describes them. As such, it is easy to envision an individual being presented with a question and that individual going through something of a matching process comparing themselves to each of the possible categories. In this case, the better the match between any given response category and an individual's self- evaluation, the higher the probability is that the individual will endorse a given response option. The choice of which response model an individual uses in order to respond to personality or attitude based questions would have a large effect on the ability levels estimated by the model. If an individual responds to questions using an ideal point response process but the data are analyzed using a dominance model, model-data misfit would probably occur and a fit

statistics' ability to detect this difference would be very important. An added benefit of the current study is that a number of previous tests of IRT model fit have been limited to constraining only the difficulty, discrimination and guessing parameters and assessing fit between different 1-, 2-, and 3-parameter logistic models. This study looks at the effect of using different response processes by analyzing data which is created via an ideal point process with fit statistics which assume a dominance response process, the two parameter logistic model.

Research Question 6: Which item fit index bests detects when the incorrect IRT model is used to analyze item data?

Adjusted Chi Square / Degrees of Freedom Ratio Test Cutoffs

In addition to comparing the Type I error and power estimates between these three fit indices, the standard cut off value for interpreting model misfit for the adj. χ^2/df ratio test will be tested. As reported by (Drasgow et al, 1995), the industry standard cut value for interpreting model fit for an adjusted χ^2/df ratio test is 3.0. This simulation study will also investigate whether or not a χ^2 to degrees of freedom ratio of 3.0 should be the standard cut score for establishing if an item is displaying model misfit.

Research Question 7: What is the best cutoff value for adj. χ^2/df ratio test?

Overall Performance

The final research question concerned the overall performance of item fit indices. That is, which test presents the best balance of Type I error rates and power?

R*esearch Question 8*: Which test presents the best balance of Type I error rates and power?

II. METHOD

Design

In order to assess the performance of the three fit indices, simulation studies were conducted in which the number of examinees simulated, the number of items on the simulated test, the number of items within a test which display a lack of local independence, the response model used for answering the questions, and the theta distribution were manipulated. A complete list of the conditions simulated is provided in Appendix B.

In order to determine the sensitivity of the three fit indices to different sample sizes and sparseness in expected and observed frequencies the sample sizes of 500, 1,000 and 2,000 were simulated. These numbers also mirror prior simulation research (Orlando & Thissen, 2003, Stone, 2003, Stone & Zhang, 2003). Three test lengths were simulated (L = 10, 20 and 40). The number of items was manipulated because this varies the precision with which ability is estimated. A test which has 10 items will have a more imprecise measurement of ability than a test which consists of 40 items. A ten-item scale was simulated to represent tests within the area of personality and attitude assessment. The use of ten items is standard in the area of psychological measurement due to relative short length and higher inter-item reliabilities compared to shorter tests. Simulation of a test with 10 items may demonstrate the difference between Stone's fit statistic, which was created for shorter tests, and the other two fit statistics. Many tests have an even larger number of items to ensure a high reliably. Tests of cognitive ability are generally slightly

longer as an artifact from development using CTT methods. 20-item and 40-item tests were also simulated in order to demonstrate how the fit indices function for these types of tests.

Local independence was manipulated by varying the number of items within the test which displayed local item dependence (LD). For each test length three levels of local independence violation were simulated. A condition which displays no LD (0%), a small amount of LD (20%) and a condition which displays a large amount of LD (40%) were estimated and tested for model fit. Two theta response distributions were estimated. A standard normal distribution, with a mean of 0.0 and a standard deviation of 1.0 and a distribution with a restricted range were estimated. The restricted range will still have a mean of 0.0, but will only have a standard deviation of 0.75. This will somewhat limit the range of thetas which are used to estimate model parameters. Finally, the effect of the using an incorrect response model was tested by analyzing data created under an ideal point methodology with a dominance IRT model. The model used to create parameter estimations will be the 2PLM for dichotomous data.

Data Generation

For all simulations data was generated via SPSS 12.0. The first step was to generate population values for the parameter estimates for 100 samples. Following the procedure used by Roberts, Donoghue, and Laughlin (2000), a parameters were randomly drawn from a continuous uniform distribution ranging from .7 to 2. B parameters were randomly drawn from a continuous uniform distribution ranging from -2 to 2. Second, a θ value for each individual simulated was sampled from a random standard normal distribution, except in the condition in which theta values were restricted. Third, the

probability of endorsing each item was calculated using the item parameters, θ values, and the appropriate IRT equations (2PLM or GGUM). Finally, item data was generated by comparing random variables with uniform distributions ranging from zero to one with the probabilities calculated. The response option for which the probability exceeded the random number was the response for that given item. Analysis of each condition was repeated 100 times.

Ideal Point Data Generation

A second response model, using an ideal point methodology, was used to introduce misfit. In order to generate ideal point response data, an ideal point methodology for data generation as opposed to the dominance method used by the two parameter logistic model, was employed. As stated above, ideal point methodology is based on the concept that non-endorsement of an item is not necessarily a function of not having a high enough level of a certain trait. The ideal point methodology is instead based on the concept that it is possible that individuals do not endorse an item because it either represents too much or too little of the given construct. If an individual believes that their trait level is less than the trait level indicated by the item it is called disagreeing from below, and individual believes that their trait level is higher than the indicated item, it is termed disagreeing from above. The latter of these two will lead to the expected item score to have a non-monotonically increasing or bell-shaped relationship with the underlying trait.

Generation of Item Parameters for Ideal Point Data

There are a number of ideal point IRT models, but in the present study, I used the generalized graded unfolding model (GGUM, Roberts, et al., 2000). This model seems to

be the most applicable to personality data and is the most general (Stark et al. 2006). The GGUM equation for dichotomous data is:

$$P_{ij}(Y=1|\theta_j) = \frac{\exp(\alpha_i[(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{i1}])}{1 + \exp(\alpha_i[3(\theta_j - \delta_i)]) + \exp(\alpha_i[(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{i1}])} \quad (14)$$

In this equation, α_i is the item discrimination parameter, θ_j is the individual j trait level, δ_i is the item location of item i on the trait level scale, and τ_i is the subjective response category threshold on that trait scale. Both α_i and τ_i determine the shape of the curve. As α_i increases, the height of the curve increases (i.e., the probability of endorsement approaches 1.0) and the peak becomes steeper. The height of the curve increases but become less steep as τ_i increases. In this model, the probability of endorsement is at a maximum when θ_j is equal to δ_i. More detail on the GGUM model is presented by Roberts et al. (2000) and Stark et al. (2006).

Following the procedure used by De Mars (2004), for each test length (10, 20, and 40) item locations (δ) will be randomly selected from a uniform distribution bounded by -3.0 and 3.0. Item discriminations (α) will be randomly selected from a uniform distribution between .5 and 2 as suggested in (Roberts, Donoghue, & Laughlin, 2000). Also following DeMars (2004) and Roberts et al. (2000), τ_c, the last τ for each item, will be randomly drawn from a uniform (-1.4, -0.4) distribution, and each preceding τ will be created by subtracting a value drawn from a random normal distribution.

Analyses

Item parameters were calibrated using Multilog 7.0 (Thissen, 2003). S-χ^2 and Stone's χ^{2*} were estimated using the SAS IRTFIT macro (Bjorner, Smith, Stone, & Sun, 2007). Adjusted χ^2/df's were computed using Microsoft Excel macros. The S-χ^2 and

Stone's χ^{2*} fit indices were calculated based on all of the cases for each sample, which is consistent with prior research (Orlando & Thissen, 2003; Stone & Zhang, 2003). As suggested by Drasgow et al. (1995), adjusted χ^2/df ratios were computed with cross-validation, and without cross validation. When cross validation was done, the item parameters were calibrated using the first half of the sample and expected frequencies were computed using the second half of the sample. Adjusted χ^2/df ratios were calculated both with and without cross-validation for this study. This allowed for a more direct comparison with the other item fit indices.

Type I error rates and empirical power estimates are displayed in individual tables for each of the simulated conditions. Type I error rates were calculated by taking the percentage of times the index indicated that an item did not fit when no manipulations or model misspecification were made. For S-χ^2 and Stone's χ^{2*} misfit was indicated when the chi-squares were statistically significant using an alpha level of .05. For the initial evaluation of misfit using the adjusted χ^2/df's, misfit was present when the value exceeded the recommended cutoff of three. Cutoff values of 1.0, 2.0, and 4.0 were also investigated.

III. RESULTS

Type I Error Rates

Table 1 presents the Type I error rates for all three of the goodness-of-fit indices for the combinations of the of test length and sample size using a standard normal distribution. An alpha level of .05 was adopted for the fit analyses for S-χ^2 and Stone's χ^{2*} and misfit is indicated for adjusted χ^2/df's above 3.0. Numbers within Table 1 represent the percentage of times across one hundred replications in which item misfit was detected in tests with no intentional introductions of misfit using a two parameter logistic model for dichotomous data. Percentages were also averaged over items. As can be seen in Table 1, Type I error rates below five percent for all sample sizes and test lengths except one were observed for S- χ^2 and Stone's χ^2*. With the exception of the adjusted χ^2/df ratio test with cross validation, all conditions had Type I error rates at or below .05. The percentage out of 100 samples and across all items which the fit indices incorrectly identified as misfitting ranged from 1 to 5, except for the adjusted χ^2/df ratio test with cross validation, but there was no discernable pattern among the conditions. The Type I error rates did not appear to lessen with either larger sample sizes, items, or the interaction between the two. Type I error rates for the adjusted χ^2/df's using cross validation were found to be unacceptably large for the conditions with both ten and twenty items. The conditions with ten or twenty items using cross validation had adjusted χ^2/df's ranging from 35% to 66%. The forty item conditions did not have this issue as

shown in Table 1, but for the sake of clarity and continuity and because of the rarity of use of cross validation due to the need for extremely large sample sizes, all further analyses, and all results are reported for the conditions without cross validation. The Type I error rates for the adjusted χ^2/df's without cross validation for both item doubles and triples displayed smaller levels of variation than the other two statistics. The Type I error rate for the adjusted χ^2/df's without cross validation varied between zero and .02 for both item doubles and triples.

Adjusting $\chi^{2'}$s that are less than their degrees of freedom always results in negative values and these negative values are normally set to zero. This occurred for all of the item singles across all the Type I error conditions and samples as can be seen in Table 1. For this reason, results based on item singles will not be considered or discussed.

All of the fit indices had smaller than expected Type I error rates, and none of the fit indices consistently had Type I error rates at the desired level of .05. Of the three indices S- χ^2 had the 'best' Type I error rates for the conditions with 10 and 20 items and Stone's χ^{2*} had the 'best' Type I error rate for the conditions with 40 items.

Effect of Sample Size

Type I error rates did not vary for S- χ^2 or Stone's χ^2* in any consistent way. As represented in Table 1, Type I errors for adjusted χ^2/df's with cross validation for both item doubles and triples tended to decrease as sample size increased. Although all three of the statistics were relatively stable, Type I errors for adjusted χ^2/df's without cross validation for both item doubles and triples were relatively the most and unaffected by sample size among all of the statistics tested.

Depending on the condition, certain fit statistics were more or less affected by sample size regarding estimates of power. As displayed in Table 2, for the condition with 20% LD, adjusted χ^2/df ratio test without cross validation for item doubles appeared to be the least affected by sample size, and for the condition with 40% LD the adjusted χ^2/df ratio test without cross validation for item triples was the least affected by sample size. None of the indices were affected in a meaningful way by limiting the range. For conditions based on analyzing data with a 2 PLM which were created under an ideal point responding process S-χ^2 appeared to be least affected by sample size. Results of the power analyses were also collapsed across conditions in order to have an overall estimate of power based on sample size. These results are presented in Table 3. Power estimates were averaged across all of the conditions which introduced misfit within a sample size in order to get an average estimate of the degree to which sample size effected the estimation of fit. Across the three conditions which introduced misfit in to the model, the adjusted χ^2/df ratio test without cross validation for item triples appeared to be the fit statistic least affected by changes in sample size, while S-χ^2 appeared to be the most affected. In conclusion, based on both Type I error rates and power estimates, the adjusted χ^2/df ratio test without cross validation is the test which is least affected by changes in sample size.

Effect of Test Length

Although none of the fit indices' Type I error rates were affected by test length to a large degree, the adjusted χ^2/df ratio test without cross validation using item triples was the least affected. This can easily be seen in Table 4. The average estimate of power across the three conditions which introduced misfit only changed by one percent across

the three different test lengths for the adjusted χ^2/df ratio test without cross validation using item triples. Adjusted χ^2/df's for item triples without cross validation was in fact totally unaffected by changes in the number of items when using a sample size of 2000 and only had changes of .01% between test size at the other sample sizes of 500 and 1000. In regards to power, the adjusted χ^2/df ratio test without cross validation for item doubles was the least affected by test length for the conditions with 20% LD. Adjusted χ^2/df's without cross validation for item triples was affected the statistic least affected by test length for the condition with 40% LD. Again, no test was affected much by any changes in test length when a restriction in range of theta values was introduced. The adjusted χ^2/df ratio test without cross validation for item doubles however, was the least affected by test length for this condition. S- χ^2 again appeared to be least affected by test length for conditions based on model misspecification based on analyzing data with a 2 PLM which were created under an ideal point responding process. In conclusion, based on the results of both the Type I error rate and power analyses, the adjusted χ^2/df ratio test without cross validation for item doubles appears to be the least affected by test length. However, the adjusted χ^2/df ratio test without cross validation for item triples could also be used without a large change in the degree to which the test is affected by test length.

Violations of Local Independence (LD)

Tables 2 presents empirical power estimates for the goodness-of-fit statistics at the combinations of the manipulated factors (test length, sample size). Adequate power is achieved for a fit index if eighty percent of the samples are identified as containing misfitting items. None of the conditions in which twenty percent of the items were created to have violations of the local independence assumption had adequate power.

Across all three of the fit indices the power estimates ranged between four and sixty eight percent out of 100 samples across all items. Specifically, the power estimates for S- χ^2 ranged between four and forty-five. Power estimates for Stone's χ^2* ranged between ten and sixty-three, and analytical power estimates for adjusted χ^2/df's ranged between zero and seven for singles, between twenty-three and forty-nine for doubles, and between twenty-seven and sixty-eight for item triples.

Although none of the conditions displayed adequate power, the pattern among these results demonstrates that in general, the conditions with one thousand simulated respondents had a larger proportion of simulated respondents displaying missfitting items than the conditions with two thousand simulated respondents. The conditions with five hundred simulated respondents identified the smallest percentage of samples displaying misfit. Furthermore, the adjusted χ^2/df test for item triples on average had the highest power ratings, and in a relative sense appeared to best detect this introduction of misfit.

40% LD

As displayed in Table 2, S- χ^2 only had one condition which displayed an adequate empirical power estimate. The power estimates for S- χ^2 ranged from eighteen to eighty-one percent. The condition with two thousand simulates and ten items obtained adequate power and all other conditions failed to obtain adequate power. None of the conditions for Stone's χ^2* obtained adequate power under this condition. The power estimates for Stone's χ^2* ranged between twenty-three to sixty-six percent. For both fit indices the percentages increased as the number of simulated respondents increased from five-hundred to two thousand. Item singles for the adjusted χ^2/df ratio without cross validation test did not display adequate power for any of the conditions. The power

estimates for item singles for the adjusted χ^2/df ratio test without cross validation ranged between four and sixteen. None of the power estimates for item doubles displayed adequate power either. Power ranged between thirty-three and seventy-seven for item doubles. Multiple conditions displayed adequate power when using the adjusted χ^2/df ratio test for item triples without cross validation to assess model-data fit. All three of the conditions for which two thousand respondents were simulated displayed adequate power. The condition with one thousand simulated respondents and forty items also achieved an adequate power estimate (.81). In response to the fourth research question, the adjusted χ^2/df ratio test for item triples without cross validation is the test which best detects violations of local independence, although none of the fit indices were able to detect this form of introduced misfit on a consistent basis.

Restriction in Range

All three of the fit indices proved to be robust against the assumption that a standard normal distribution needs to be used for the estimation of possible theta values when estimating the IRT model. This manipulation proved to be not so much an estimation of power, as a display of the robustness of the statistics. Narrowing the range of possible theta values from a range between 0 and 1.0 to a range of 0 to .75 did not appear to negatively affect any of the fit statistics. All of the fit indices displayed not only adequate Type I error rates, but generally identified a fewer number of misfitting tests. The results for the range restricted conditions were similar to the Type I error results. Both S- χ^2 and Stone's χ^2* maintained Type I error rates between one and five percent. Type I error rates for the adjusted χ^2/df ratio test without cross validation for singles, doubles and triples all ranged between zero and three. Much like the Type I error

conditions, no pattern of results emerged among the various conditions. In conclusion and in response to the fifth research question, although misfit was not introduced by restricting the range of possible theta values, the adjusted χ^2/df ratio test without cross validation for doubles and triples was relatively the least affected by this change compared to the other two fit statistics.

Model Misspecification with the GGUM

The last type of misfit introduced into the model was the use of the GGUM to generate data. After generating the data with the GGUM, the data was then analyzed with a two parameter logistic model. Table 2 also displays the results for the GGUM condition. Under this condition S- χ^2 did not obtain adequate power across any of the nine combinations of item number and sample size. The power estimates for S- χ^2 ranged between fourteen and forty-one. No consistent pattern of results among the nine conditions for S- χ^2 was found. The power estimates for Stone's χ^2* ranged between forty-one and ninety percent of the samples exhibiting misfitting items. Adequate power was found for two of the conditions for Stone's χ^2*. The twenty item condition had adequate power for one thousand and two thousand simulated respondent conditions. The power estimates were .83 and .90 respectively. The adjusted χ^2/df ratio test without cross validation did not achieve adequate power for any of the conditions across all of item singles, doubles and triples. The fit statistic which appears to best detect misfit based on model misspecification stemming from a situation in which data is analyzed with a two parameter logistic model, but was created under an ideal point methodology is Stone's χ^{2*}.

Type I Error Rates for Different Cut-Off Values

Table 5 displays the Type I error rates for the adjusted χ^2/df's without cross validation at the cut-off value of 1.0. When reducing the cut-off value for adjusted χ^2/df's without cross validation from the standard value of 3.0 to a lower number, the test becomes increasingly conservative. The test for singles under this situation had Type I error rates of zero across all conditions. The cut-off value of 1.0 proved to be too conservative for some, but not all of the conditions for doubles and triples. Using 1.0 as a cut-off value, the adjusted χ^2/df ratio test without cross validation did not have an acceptable Type I error rate for four out of nine conditions for doubles and six out of nine conditions for triples.

Table 6 displays the Type I error rates for the adjusted χ^2/df's without cross validation at the cut-off value of 2.0. Adjusted χ^2/df's without cross validation for singles again resulted in all zeros. The doubles and triples for the cut-off value of 2.0 however displayed acceptable Type I error rates across all of the eighteen conditions between doubles and triples. The percent of samples identified as demonstrating misfit ranged between one and four for doubles and between zero and four for triples.

Raising the cut-off value from 3.0 to 4.0 also caused the statistic to report unacceptable Type I error rates. As can be seen in Table 7, raising the cut-off value from 3.0 to 4.0 resulted in a higher percent of conditions displaying unacceptably low Type I error rates. The singles condition had a Type I error rate of zero for all conditions. Type I error rates of zero were found in all of the one thousand and two thousand simulated respondent conditions for doubles and triples. The adjusted χ^2/df ratio test had Type I

error rates of one for both doubles and triples for all three of the conditions with five hundred simulated respondents.

Power Estimates for Different Cut-off Values

Cut off Value of 1.0

In order to gauge the adjusted χ^2/df ratio test without cross validation's ability to detect misfit at the various cut-off values and determine if either a higher or lower cut-off value would work as effectively or even better, a number of power analyses were run at the different cut-off values. Table 8 shows that at the cut-off value of 1.0, the adjusted χ^2/df ratio test for item triples had sufficient power to detect the introduction of twenty percent LD misfit for conditions with both one thousand and two thousand simulated respondents. The adjusted χ^2/df ratio test for item triples did not have adequate power for the conditions with only five hundred simulated respondents. No conditions for either item singles or item doubles achieved an adequate power estimate. Table 8 also shows the 40% LD condition. The adjusted χ^2/df ratio test without cross validation did not have adequate power for item singles. Two of the conditions for item doubles achieved adequate power. The i=10 & 40 conditions with N= 2000 had adequate power. All other conditions for item doubles failed to achieve adequate power. When forty percent of the items were created with a violation of the local independence assumption for the adjusted χ^2/df ratio test without cross validation for item triples, only two conditions failed to achieve an adequate power estimate. The two conditions which did not have adequate power were two conditions with five hundred simulated respondents for both ten and twenty items.

Using a lowered cut-off value of 1.0 instead of 3.0 for the adjusted χ^2/df ratio test without cross validation when testing data in which a restricted range was used created a situation in which the fit statistic reported higher rates of misfit than when a cut-off value of 3.0 was used when assessing the statistic using item doubles or item triples. Table 8 shows that item singles still ranged from zero to one, but item doubles ranged from three to five and item triples ranged from five to nine. None of the conditions with the GGUM generated items achieved adequate power estimates. All of the conditions for singles, doubles and triples ranged between one and forty-nine.

Cut off Value of 2.0

Lowering the cut-off value to 2.0 led to more conditions which achieved adequate analytical power estimates compared to a cut off value of 3.0, but fewer conditions which achieved acceptable power estimates than a cut off value of 1.0 for the adjusted χ^2/df ratio test without cross validation. As can be seen in Table 9, when twenty percent of the items were created to violate the assumption of local independence none of the conditions achieved adequate power. When forty percent of the items were created to have a violation of the local independence assumption, none of the conditions for item singles achieved adequate power and only one condition, forty items and two thousand simulated respondents, for item doubles achieved adequate power. For item triples none of the conditions with five hundred simulated respondents achieved adequate power. The condition with forty items and one thousand simulated respondents achieved adequate power. All three of the conditions for ten, twenty and forty items achieved an adequate power estimate at two thousand simulated respondents.

Lowering the cut-off value to 2.0 instead of 1.0 for the adjusted χ^2/df ratio test without cross validation when using data with a restricted range created results which were similar to the standard cut-off value of 3.0. None of the conditions achieved acceptable power. Table 9 shows that item singles ranged from zero to one, doubles and triples both ranged between one and four. The GGUM generated items tested with an adjusted χ^2/df ratio test without cross validation with a cut-off value of 2.0 showed the same general pattern as it did when the cut-off value was lowered to 1.0. All of the conditions failed to achieve adequate power.

Cut off Value of 4.0

Raising the cut-off value for acceptable fit from 3.0 to 4.0 made the adjusted χ^2/df ratio test without cross validation somewhat more conservative. The results for these conditions are summarized in Table 10. Testing the conditions in which twenty percent of the items were created with a violation of the local independence assumption with the adjusted χ^2/df ratio test without cross validation at this higher cut-off value led to similar results as the other cut-off values. None of the conditions displayed adequate analytical power estimates. The power estimates for item singles ranged between zero and six. The power estimates for item doubles and triples ranged from thirteen to forty-two percent and between twenty-two and sixty for triples respectively. For this set of conditions, the condition with one thousand simulated respondents again generally had higher percentages than the condition with two thousand simulated respondents and the condition with five hundred simulated respondents had the lowest reported power estimates.

As displayed in Table 10, when the cut-off value was set to 4.0 none of the conditions had adequate power estimates for item singles or item doubles when the test was run on samples in which forty percent of the items were created to have a violation of the local independence assumption. For item triples each of the conditions with two thousand simulated respondents achieved adequate power and all other conditions for five hundred and one thousand simulated respondents failed to achieve adequate power. The general pattern for results in this condition showed that as the number of simulated respondents and the number of items increased, a larger percentage of samples displayed misfit according to the adjusted χ^2/df ratio test without cross validation.

The conditions with a restriction in range of theta values not surprisingly all had very low percentages times of reported misfit and failed to achieve adequate power. The values ranged from zero to five across all singles, doubles and triples. Again, the GGUM generated items continued to follow the same pattern in which none of the conditions displayed adequate power estimates. In regards to Research Question 7, it appears that the standard cut-off value of 3.0 may not be the ideal cut-off value; at least for the conditions assessed in the current study. A cut-off value of 2.0 had acceptable Type I error rates, while still maintaining the same level of power as a cut-off value of 3.0.

The final Research Question concerned which of the fit statistics had the best trade-off between Type I error rates and power. As stated earlier, the assumed trade-off between Type I error and power did not appear to occur. Although the adjusted χ^2/df ratio test had the lowest Type I error rate out of the three fit indices, it also had the highest power estimates for 2 out of the 3 conditions which actually did introduce misfit into the

model. As such the adjusted χ^2/df ratio test appeared to have both the best Type I error rate and power.

IV. DISCUSSION

Summary Results

The present study had four main goals. The first goal was to assess the different Type I error rates for the fit indices at different levels of sample size and items. The second goal was to determine power estimates for each of the fit statistics under a number of different conditions which introduced misfit. This included the fit indices' ability to detect misfit stemming from analyzing data with an IRT logistic model when the data had been generated under an ideal point situation. This type of misfit had not been evaluated in any prior simulation study, and was of the most interest to the current study. The third goal was to critically analyze the standard cut-off value used for the adjusted χ^2/df ratio test, another question which to the knowledge of the author had never been empirically tested. This was done for the adjusted χ^2/df ratio test without cross validation due to high Type I error rates caused when cross validation was used. The final goal was to identify the 'best' fit index out of the three analyzed. The first two goals were meant to give an indication of which index should be used given the conditions considered in the present study. The third goal was intended to be an empirical assessment of the suggestion posed by Drasgow, et al (1995) that the preferred cut-off value for the adjusted χ^2/df ratio test should be 3.0. And the final goal was to come up with a suggested test based on all of the relevant factors which may come into play when the evaluation of fit of an IRT model is to be undertaken.

The results in Table 1 show that all three of S- χ^2, Stone's χ^{2*} and the adjusted χ^2/df ratio test without cross validation the exhibited lower than expected Type I error rates. The Type I error rates were somewhat lower than expected in that many of the conditions had Type I error rates below .05. Although none of the fit indices displayed a clear ability to detect the misfit at an acceptable level across all conditions, certain tests performed better than others under certain conditions. Furthermore, adjusted χ^2/df's without cross validation exhibited low Type I error rates, however the adjusted χ^2/df's with cross validation resulted in far too many Type I errors. The high Type I error rates found in the present study for the adjusted χ^2/df ratio test with cross validation for dichotomous data suggest that using cross validation is not suggested for the detection of misfit for tests with either ten or twenty items when sample sizes even as large as 2,000 are used. The difference in Type I error rates for the index when using cross validation between twenty and forty items may possibly be accounted for by the increase in variation in the parameters estimated by the IRT model. This variation may have compensated low sample sizes used. Although all three of the fit indices displayed lower than expected Type I error rates, the adjusted χ^2/df's without cross validation also displayed the highest power estimates and as such suggests that the lower than expected Type I error rates are not a hindrance for the power of the statistic.

None of the fit indices consistently had adequate power for the conditions with lower levels of LD. When forty percent of the items were created to have items which were locally dependent S-χ^2 displayed adequate power for the condition with ten items and two thousand simulated respondents, but no other conditions. The adjusted χ^2/df ratio test with cross validation for item triples displayed adequate power for the most number

of conditions when considering a test with 40% LD. Stone's χ^2* did not have any conditions which displayed adequate power for any of the LD conditions. Stone's χ^2* however had the best analytical power estimates for the model misfit condition in which data was created under a GGUM and analyzed with a 2PLM; Stone's χ^2* had two conditions which displayed adequate power. The only other conditions which displayed adequate power was the adjusted χ^2/df ratio test with cross validation, which only identified one condition across all three of item singles, double and triples.

No pattern of expected results was predicted, but in general it would be desirable for each of the statistics to not be affected by either sample size or test length. This did not occur. Both sample size and test length affected power estimates. Power was also affected as sample size and or the number of items increased for the LD conditions. Further, the restriction of range generally did not appear to affect power estimates across either sample sizes or test lengths. All three fit indices maintained low Type I error rates when the range of possible theta values was restricted from zero to one to a smaller range of zero to .75. Sample size and test length also affected power estimates when misfit was introduced based on model misspecification. The results suggest that overall, the adjusted χ^2/df test was the least affected by changes to sample size and test length.

Research Questions

Research question 1 was concerned with which statistic had the best Type I error rate. As can be seen in Table 1, no consistent pattern emerged for Type I error rates among any of the fit indices. The adjusted χ^2/df however had the lowest Type I error rate while also maintaining the best estimates of power. The use of the statistic without cross validation goes against suggestions by some that cross validation is a necessary step in

order to ensure a psychometrically sound statistic. This is because not doing so can lead to a representation of the fit of the data to the model which is inflated for certain fit statistics such as the adjusted χ^2/df ratio test. Despite this fact, in many situations a cross validation sample is not used due to the very large sample sizes needed to cross validate the data. Research questions 2 and 3 inquired about how each of the fit indices' Type I error rates would be affected to changes in sample size and test length respectively. As shown in Tables 3 and 4, the statistic least affected by either of these two changes was the adjusted χ^2/df ratio test without cross validation. Sample size least affected adjusted χ^2/df ratio test without cross validation when item triples was used. The adjusted χ^2/df was least affected by test length when item doubles was used. The use of multiple items to calculate the statistic is a possible reason why this statistic was affected by these changes to a smaller degree than the other two statistics. The combining of items may have mitigated some of the affect variation in the estimation of item parameters.

Research question 4 asked which fit statistic would be best able to detect misfit caused by a violation of the local independence assumption. For the condition with twenty percent LD, power estimates did not increase as sample size increased. For the conditions with twenty percent LD and across all three fit indices, the condition with one thousand simulated respondents tended to have the higher analytical power estimates than the other two conditions. For the conditions with twenty percent LD, as the number of items increased, empirical power tended to decrease. This was true across all three of the fit indices. For the conditions with forty percent LD empirical power estimates tended to increase as the number of simulated respondents increased, and empirical power

estimates decreased as the number of items simulated increased across all three of the fit indices.

It is possible that a sample size of five hundred is not large enough to detect misfit when only twenty percent of the items have an issue with local independence, but when a sample size as large as two thousand is used, this low level of misfit is masked by the large number of observations. The increase in power estimates as sample size increased for the condition with forty percent LD may be because having forty percent of the items with a violation of the LD assumption is a large enough of a percent of the items with issues of LD for the statistics to have better estimates of power due to the increased precision the statistic obtains in the estimation of item parameters as sample size increases.

Research question 5 was concerned with how data sets which had a restriction in range would display would be affected. Restricting the range of possible thetas did not increase Type I error rates. The opposite result was found. Restricting the range of possible theta values for the IRT model to use in the estimation of the item parameters tended to decrease the number of Type I errors among all three of the fit statistics. This may have occurred due to the decrease in variability in theta values caused by decreasing the range of possible values. With a smaller range of possible theta values, there is a smaller probability that an extreme value can occur which would cause the model to be identified as not fitting.

Research question 6 inquired about empirical power estimates when data which was generated using an ideal point IRT methodology was analyzed using a dominance IRT model. This type of situation could occur if an individual responding to a question is

not does not use a monotonically increasing response pattern. If there is a point at which an individual may say to themselves "This question represents too much of this trait for me to agree with it," then using a 2 PLM to analyze the data would be an incorrect model. As can be seen in Table 2, power estimates for S-χ^2 under this specific condition tended to increase as sample size increased, but no consistent pattern emerged for test length. For Stone's χ^2* no consistent pattern regarding either sample size or test length emerged. For the adjusted χ^2/df ratio test without cross validation power tended to decrease as sample size increased, but no pattern emerged for the test length.

It is interesting that power increased as sample size increased for S-χ^2 but decreased as sample size increased for the adjusted χ^2/df ratio test without cross validation. The increase in power as sample size increased for S-χ^2 is a relatively common phenomenon. The estimation of power, or the ability to detect misfit when it truly exists, often increases as sample size increases. The contrary finding of power estimates decreasing as sample size increased for the adjusted χ^2/df ratio test without cross validation may be related to the use of item doubles and triples for the creation of adjusted χ^2/df's. The power did not decrease as sample size increased for the adjusted χ^2/df ratio test without cross validation for item singles. Stone's χ^{2*} having the best power estimates for this condition may be due to the focus on observed values in the estimation of the statistic. Spreading the pseudo-observed values across multiple theta levels may have given the statistic more chances to capture this sort of model misfit.

Based on these results, although all three of the fit indices have acceptable Type I error rates, different indices appear to be more appropriate depending on the type of misfit the individual researcher or practitioner is interested in testing for. Although none

of the fit indices were able to adequately detect misfit on a consistent basis for any of the conditions, for the conditions with LD, the adjusted χ^2/df ratio test without cross validation for item triples consistently had the highest power. The adjusted χ^2/df ratio test without cross validation for item triples not only consistently had the highest power under the conditions with forty percent LD, but also had a number of conditions which did achieve adequate power. Although the adjusted χ^2/df ratio test without cross validation for item triples had the best power estimates, it is highly suggested that item singles for the adjusted χ^2/df ratio test without cross validation not be used due to the issue with adjusting $\chi^{2'}$s that are less than their degrees of freedom. Item doubles also failed to achieve the same levels of power as item triples did. Due to its performance on the conditions concerning LD, the adjusted χ^2/df ratio test without cross validation for item triples appears to be the best index out of the three for detecting misfit based on LD. If there is reason to believe that the data being used has an issue with violating the LD assumption, the adjusted χ^2/df ratio test without cross validation for item triples would be the recommended fit statistic to use.

The adjusted χ^2/df ratio test without cross validation is also the recommended due to its' ease of use and general accessibility compared to the other two fit statistics. Further, the concepts underlying the adjusted χ^2/df ratio test are in general much simpler to understand. It is noteworthy to mention that when assessing fit with a statistic based on a chi-square distribution it is important to ensure that the samples being tested approximate a chi-square distribution. The present study found relatively small Type I error rates, which suggests that the distributions for these fit indices did approximate a chi-square distribution.

If there is reason to believe that the responses given to an assessment were created under an ideal point methodology, the best course of action would be to analyze the data using the GGUM or another IRT model which can account for this type of non-monotonic response pattern. If this is not possible and analysis must proceed with a 2PLM, the fit statistic best suited to detect this situation is Stone's χ^2*. This index consistently had the highest power estimates and had the most number of conditions with adequate power.

Since all three of the fit indices had roughly the same Type I error rates, the choice among them would come down to their ability to detect the misfit introduced. The adjusted χ^2/df ratio without cross validation and Stone's χ^2* were the two indices which performed the best under one or another of the misfit conditions. When compared against one another, the adjusted χ^2/df ratio without cross validation and Stone's χ^2* have advantages and disadvantages. One of the major disadvantages of the adjusted χ^2/df ratio test is that it is not based on a single item. The necessity of using item triples; due to item singles' inability to capture misidentification complicates the identification of items which may be the cause of misfit. It is a common practice to remove items which demonstrate misfit, and if misfit is identified by an item triple, simply removing the item that is displaying the misfit is not easy due to the fact that the same item may be in multiple combinations of triples and only one combination may indicate poor fit. Stone's χ^2* does not have the same issue with the inability to identify model misspecification as it is an index based on a single item. The adjusted χ^2/df ratio test without cross validation is however a well known, accessible and often used statistic. It is readily available free online, is very easy to compute and its' use is well documented in IRT literature.

The ability to detect issues of LD is important. This can be a common problem for many types of test in which information given or gleamed from certain items may increase the likelihood of a positive response on one or even sometimes a number of other items. The ability to detect when the GGUM was used to generate the data was displayed for Stone's χ^{2*}, but although this is an important problem, it may not be as common as issues with local dependence. The ability of a fit index to make this distinction is important however, because recently there has been some question as to whether individuals may respond to personality items using an ideal point rather than a dominance response process. (Chernyshenko, et al 2007).

S- χ^2 and Stone's χ^{2*} had Type I error results consistent with previous research. Previous simulation research found these fit indices to have nominal Type I error rates for several dichotomous IRT models (Stone & Zhang, 2003). Rejection rates for S- χ^2 were found to be close to .05 for tests with ten, forty and even eighty items. Stone's χ^{2*} also achieved acceptable Type I error rates for twelve items, but not for six items at *n*'s of five hundred, one thousand and two thousand.

When using cross validation the adj. χ^2/df ratio test fit statistic reported unacceptable Type I error rates. This may be caused by the smaller sample sizes used in the simulation, and emphasizes the need for larger sample sizes when estimating IRT models. The results of this study also suggest that the use tests with a larger number of items may mitigate this problem. When assessing samples with forty items, the problem of unacceptably high Type I error rates disappeared completely.

The use of cross validation in analysis depends on what the reason for the analysis is, the goals of the research question being asked, and the sample size collected. Not all

item fit indices utilize cross validation. Studies which did not use cross validation tended to be an investigation of the aspects of, or differences within, a specific sample. Examples of this are investigations of DIF or research questions asking the appropriateness of using IRT modeling for a type of item or sample. The use of cross validation on samples appears to apply more to research questions that will use the item parameters estimated in future situations or samples, such as the development of scales.

Research question seven asked whether or not the standard value of 3.0 as a cut off value for the adjusted χ^2/df test is a good standard to use to identify misfit. No alternative cut off value was suggested, but three alternative cut off values were investigated. Alternative cut-off values of 1.0, 2.0 and 4.0 were investigated. The results indicated that a cut-off value of 1.0 was too strict of a value and resulted in too many Type I errors. Both cut-off values of 2.0 and 4.0 had acceptable Type I error rates.

Tables 6 through 8 shows the results of the power analyzes at the different cut off points. Although a cut off value of 1.0 appears to have the highest power estimates, the cut off value of 1.0 also had unacceptable Type I error rates. Setting the cut off value at 4.0 led to acceptable Type I error rates, but lower power estimates than setting the cut off value at 3.0. Setting the cut off value at 2.0 appears to have Type I error rates and power estimates comparable to the cut off value of 3.0. For the cut off value of 2.0, all of the conditions had acceptable Type I error rates for the adjusted χ^2/df ratio test without cross validation for item triples. The cut off value of 2.0 also had the highest number of conditions with adequate power estimates among all of the alternative cut off values, including the value of 3.0 often used.

The cut-off value of 2.0 appears to be neither too conservative nor liberal. While maintaining a good Type I error rate, the adjusted χ^2/df test using a cut-off value of 2.0 was still able to identify misfit. This suggests that perhaps the convention of using a cut-off value of 3.0 is perhaps unnecessarily liberal. If a more conservative test is able to perform equally well as a more liberal test, it would be a benefit to use the more conservative test and be more confident in the findings of the fit statistic.

Implications

Given the results of this study, overall, the fit index which appears to in general be the best for analyzing the fit of a two parameter logistic model with dichotomous data is the adjusted χ^2/df ratio test without cross validation for item triples. The adjusted χ^2/df ratio test without cross validation had the best Type I error while maintaining the best power out of the three fit statistics investigated. This fit statistic is also recommended due to ease of use and the accessibility The adjusted χ^2/df ratio test however may not be as useful for the detection of misfit at the individual item level. When investigating the fit of individual items, it may be helpful to use Stone's χ^2*, as it is a fit statistic based on single items.

The present study also demonstrated that given very specific situations, different fit statistics may be better at identifying different types of misfit. The adjusted χ^2/df ratio test without cross validation appears to be best at detecting misfit based on issues with LD. In order to detect misfit based on an incorrect model specification between a two parameter logistic model and an ideal point responding process, Stone's χ^2* appears to be best suited for the detection of misfit. This is important because it has been suggested that individuals may answer certain personality type questions with an ideal point

methodology in mind. If a model fails to fit when assessed under a dominance model, it may be that the questions were in actuality answered using a different model, and a different IRT model may be more appropriate for the assessment of fit. The recommendation would be to use the adjusted χ^2/df ratio test without cross validation as a tool for assessing overall model-data fit, and using Stone's χ^2* if there is reason to believe model misspecification is a possible issue.

Furthermore, when using the adjusted χ^2/df ratio test with cross validation, this simulation study demonstrated that it may be possible to use an adjusted χ^2/df ratio of 2.0 as the cut-off value for determining acceptable fit instead of 3.0. Changing the cut-off value from 3.0 to 2.0 may not work for all conditions, but for the specific conditions simulated in this study, the index appears to maintain both acceptable Type I error rates and adequate power. The current findings are an encouraging display of the fit index's ability to perform at a different cut-off value.

Further, the results of the present study in combination with the results of Stone (2000), Stone and Zhang (2003) and Drasgow et al. (1995) demonstrate the capability of all three of the fit indices to detect a number of different types of misfit. The majority of types of misfit introduced in prior research were based on the use of non-model fitting conditions in which the model used to simulate data was different from the model used to calibrate the data (Stone & Zhang, 2003) by either adding or subtracting the number of parameters estimated. The present study instead actively created misfit by manipulating item characteristics such as theta range and local independence. The one condition which did utilize non-model fit did so with a model (GGUM) which has not received as much attention in the IRT fit literature as logistic models.

Limitations and Future Research

The present study is not without limitations. A primary limitation of this study is that the inclusions of all of the possible causes of misfit were not incorporated. Included in this study were manipulations of sample sizes and test lengths which allowed for direct comparisons with previous research. This study manipulated only certain item characteristics. The present study did not manipulate some item characteristics which have been manipulated in other studies. Two such manipulations that were not included were the manipulation of the difficulty parameter and or the discrimination parameter. For example, Stone and Zhang (2003) simulated item responses with a slope parameter of 1.2 and the fit statistic was computed with the slope of 0.7. The exclusion of this type of manipulation may have had an effect on the conclusions made.

Consistent with previous research, the lowest sample size was 500 and the largest sample size was 2000 cases. The sample size of five hundred is regularly used to establish the sensitivity of the fit statistics to sparseness in observed and expected frequencies. This study did not examine samples sizes outside of the conventions set forth by prior examinations of goodness of fit. It is possible that the fit statistics would still function well with a lower sample size. Knowledge that IRT modeling can be successfully achieved with lower sample sizes would be very useful, especially for applied settings when sample sizes as large as five-hundred are difficult to achieve.

Further research into the manipulation of theta values is also needed. This study found that when the range of possible theta values was restricted there was not a negative effect on Type I error rates. It was expected that restricting theta would cause the fit indices to identify more items as misspecifed. As this was not the case, it would be

interesting to assess the fit indices' ability to detect misfit with both a restricted theta and with LD issues.

Finally, research should continue on assessing a more specific and empirically proven cut-off value for the adjusted χ^2/df ratio test. The present study suggested that a number of other possible cut-off values are possible, but it may be that under certain conditions these alternative cut-off values would not work as well. It is also possible that cut-off values at non-integer values such as 1.5 would prove to be a more precise assessment of an acceptable cut-off value for this statistic.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. Kit, Trans.). Baltimore, MD:Williams & Wilkins.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems.* Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas: January.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability*. Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1958b). *On the estimation of mental ability*. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas: January.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16,* 21-33.

Bjorner, J.B, Smith, K.J., Stone, C.A., & Sun, X. (2007). IRTFIT: A macro for item fit and local dependence tests under IRT models. Lincoln, RI: Quality Metric, Inc.

Bolt, D.M., Hare, R.D., Vitale, J.E., & Newman, J.P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment, 16,* 155-168.

Chernyshenko, O.S., Stark, S., Chan, K.Y., Drasgow, F., & Williams, B.A.

(2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523-562.

Chernyshenko, O.S., Stark, S., Drasgow, F., & Roberts, B.W. (2007). Constructing personality scales under the assumptions of an ideal point response process. *Psychological Assessment, 19*, 88-106.

DeMars, C.E. (2004). Type I error rates for generalized graded unfolding model indices. *Applied Psychological Measurement, 28,* 48-71.

Drasgow, Fl, Levine, M.V., Tsien, S., Williams, B.A., & Mead, A.D. (1995). Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement, 19*, 143-165.

Ellis, B.B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74(6),* 912-921.

Embretson, S. E. & Reise, S. P. (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fischer, G.H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Glas, C.A.W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*, 45-52.

Glas, C.A.W. & Hendrawan, I. (2005). Testing linear models for ability parameters in item response models. *Multivariate Behavioral Research, 40*, 25-51.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory* (Vol. 2). Newbury Park, CA: Sage.

Han, K.T. and Hambleton, R.K. (2007). User's Manual for WinGen: Windows Software that Generates IRT Model Parameters and Item Responses. Center for Educational Assessment Research Report No. 642, University of Massachusetts.

Hu, L. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to unparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Irwin.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. *Methodology in the social sciences*. New York: Guilford Press.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61*, 273–287.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-53.

Lord, F. (1980). *Applications of item response theory to testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F.N., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maydeu-Olivares, A., Hernandez , A., & McDonald, R.P. (2006). A multidimensional ideal point item response theory model for binary data**.** *Multivariate Behavioral Research, 41,* 445-471.

Masters, G.N., & Wright, B.D. (1996). The partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*, 99-114.

Muraki, E. (1997). A generalized partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reise, S.P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in item response theory. *Applied Psychological Measurement, 14*, 127-137.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. Psychometrika, 1, 33-49.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3-32.

Roberts, J.S. & Laughlin, J.E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Rost, J. & Carstensen, C.H. (2002). Multidimensional rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*, 42-56.

Samejima, F. (1969). The graded response model . In W.J. van der Linden & Hambleton, R.K. (Eds.), *Handbook of modern item response theory*. New York: Springer.

Scherbaum, Cohen-Charash, & Kern (2006). *Educational and Psychological Measurement, 66*, 1047-1063.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*, 271-295.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2005). An IRT approach to constructing

and scoring pairwise preference items involving stimuli on different dimensions: The Multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29,* 184-203.

Stark, S., Chernyshenko, O.S., Drasgow, F., Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91,* 25-39.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 91,* 1292-1306.

Stone, C.A. (2000). Monte carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58-75.

Stone, C.A. (2003). Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. *Educational and Psychological Measurement, 63,* 566-583.

Stone, C.A. & Hansen, M.A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement, 60,* 974-991.

te Marvelde J.M., Glas, C.A.W., Van Landeghem, G., Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66,* 5-34.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology,* 16, 433–451.

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychmetrika, 11*, 1-13.

Van den Wollenberg, A.L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6*, 83-91.

Yen, W. M. (1981). Using simulation results choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Zickar, M.J., Russell, S.S., Smith, C.S., Bohle, P., & Tilley, A.J. (2002). Evaluating two morningness scales with item response theory. *Personality and Individual Differences, 33*, 11-24.

Appendix A.

Differences Between CTT & IRT

CTT	
1.	The standard error of measurement applies to all scores in a particular population.
2.	Longer tests are more reliable than shorter tests.
3.	Comparing test scores across multiple forms is optimal when the forms are parallel.
4.	Unbiased estimates of item properties depends on having representative samples.
5.	Test scores obtain meaning by comparing their position in a norm group.
6.	Interval scale properties are achieved by obtaining normal score distributions.
7.	Mixed item formats leads to unbalanced impact on test total scores.
8.	Change scores can not be meaningfully compared when initial score levels differ.
9.	Factor analysis on binary items produces artifacts rather than factors.
10.	Item stimulus features are unimportant compared to psychometric properties.
IRT	
1.	The standard error of measurement differs across scores, but generalizes across populations.
2.	Shorter tests can be more reliable than longer tests.
3.	Comparing test scores across multiple forms is optimal when test difficulty levels vary between individuals.
4.	Unbiased estimates of item properties may be obtained from unrepresentative samples.
5.	Test scores have meaning when they are compared for distance from items.
6.	Interval scale properties are achieved by applying justifiable measurement models.
7.	Mixed item formats can yield optimal test scores.
8.	Change scores can be meaningfully compared when initial score levels differ.
9.	Factor analysis on raw item data yields a full information factor analysis.
10.	Item stimulus features can be directly related to psychometric properties.

*Taken directly from Embretson and Reise, 2000

Appendix B.

Conditions Simulated

Simulation Number	Number of items	Sample Size	No. of items with Dependence	Distribution	Model
1	10	500	0	SN	2PLM
2	10	1000	0	SN	2PLM
3	10	2000	0	SN	2PLM
4	10	500	2	SN	2PLM
5	10	1000	2	SN	2PLM
6	10	2000	2	SN	2PLM
7	10	500	4	SN	2PLM
8	10	1000	4	SN	2PLM
9	10	2000	4	SN	2PLM
10	20	500	0	SN	2PLM
11	20	1000	0	SN	2PLM
12	20	2000	0	SN	2PLM
13	20	500	4	SN	2PLM
14	20	1000	4	SN	2PLM
15	20	2000	4	SN	2PLM
16	20	500	8	SN	2PLM
17	20	1000	8	SN	2PLM
18	20	2000	8	SN	2PLM
19	40	500	0	SN	2PLM
20	40	1000	0	SN	2PLM
21	40	2000	0	SN	2PLM
22	40	500	8	SN	2PLM
23	40	1000	8	SN	2PLM
24	40	2000	8	SN	2PLM
25	40	500	16	SN	2PLM
26	40	1000	16	SN	2PLM
27	40	2000	16	SN	2PLM
28	10	500	0	R	2PLM
29	10	1000	0	R	2PLM
30	10	2000	0	R	2PLM
31	20	500	0	R	2PLM
32	20	1000	0	R	2PLM
33	20	2000	0	R	2PLM
34	40	500	0	R	2PLM
35	40	1000	0	R	2PLM
36	40	2000	0	R	2PLM
37	10	500	-	-	I.P.
38	10	1000	-	-	I.P.
39	10	2000	-	-	I.P.
40	20	500	-	-	I.P.

41	20	1000	-	-	I.P.
42	20	2000	-	-	I.P.
43	40	500	-	-	I.P.
44	40	1000	-	-	I.P.
45	40	2000	-	-	I.P.

Note: SN= Standard Normal, R = Restricted, 2PLM = Two Parameter Logistic Model, I.P. = Ideal Point.

Table 1

Overall Type I Error percentages for the fit indices.

			Adj χ^2/df (3.0)		
Sample Size	S-χ^2	Stone χ^{2*}	Singles	Doubles	Triples
		10 Items			
500	2	2	42 (0)	41(1)	59(1)
1000	4	3	41 (0)	52(1)	55(1)
2000	3	2	37 (0)	45(0)	44(0)
		20 Items			
500	3	3	43 (0)	59(2)	66(2)
1000	3	1	40 (0)	50(1)	50(0)
2000	4	3	35 (0)	49(0)	42(0)
		40 Items			
500	1	5	0 (0)	3(2)	3(2)
1000	2	2	0 (0)	2(1)	2(1)
2000	2	4	0 (0)	2(1)	1(0)

Note. Percentages out of 100 samples and across all items.
Values in parenthesis represent Adj χ^2/df without cross validation.

Table 2
Overall power estimates for the fit indices.

Number of items	Sample Size	S-χ^2	Stone χ^2*	Adj χ^2/df (3.0)		
				Singles	Doubles	Triples
20% items w/ LD						
10 items	N = 500	15	30	7	26	40
	N = 1000	45	61	6	44	66
	N = 2000	44	57	3	49	68
20 items	N = 500	9	17	0	16	27
	N = 1000	32	63	3	48	68
	N = 2000	25	45	1	41	56
40 items	N = 500	4	10	0	23	30
	N = 1000	23	52	3	49	68
	N = 2000	18	26	0	36	43
40% items w/ LD						
10 items	N = 500	32	54	16	46	69
	N = 1000	53	50	14	42	65
	N = 2000	81	66	8	64	89
20 items	N = 500	20	32	6	33	63
	N = 1000	46	51	12	41	67
	N = 2000	62	62	13	61	89
40 items	N = 500	18	23	4	52	73
	N = 1000	38	43	11	60	81
	N = 2000	45	57	9	77	95
Items w/ restricted θ range						
10 items	N = 500	4	5	0	2	1
	N = 1000	3	5	0	0	0
	N = 2000	5	4	0	0	0
20 items	N = 500	4	4	1	3	3
	N = 1000	4	2	0	1	1
	N = 2000	3	2	0	0	0
40 items	N = 500	2	2	0	2	2
	N = 1000	2	2	0	1	1
	N = 2000	2	2	0	0	0
GGUM Generated						
10 items	N = 500	21	45	3	14	19

	N = 1000	23	43	1	8	9
	N = 2000	27	41	0	4	4
20 items	N = 500	20	68	2	16	24
	N = 1000	29	83	2	16	22
	N = 2000	41	90	3	15	19
40 items	N = 500	14	35	4	13	29
	N = 1000	17	49	2	10	14
	N = 2000	23	53	1	7	8

Note. Percentages out of 100 samples and across all misfitting items. GGUM is the generalized graded unfolding model. Adj χ^2/df were computed without cross validating.

Table 3

Average power estimates across all misfit conditions based on sample size

No. of items	S-χ^2	Stone χ^2*	Adj χ^2/df (3.0)		
			Singles	Doubles	Triples
500	17	35	5	27	42
1000	34	55	6	35	51
2000	41	55	4	39	52

Note. Numbers represent average misfit identified across the 20%, 40% LD and GGUM generated items.

Table 4

Average power estimates across all misfit conditions based on test length

Sample Size	S-χ^2	Stone χ^2*	Adj χ^2/df (3.0)		
			Singles	Doubles	Triples
10	38	50	6	33	48
20	32	57	5	32	48
40	22	39	4	36	49

Note. Numbers represent average misfit identified across the 20%, 40% LD and GGUM generated items.

Table 5

Overall Type I Error percentages for the Adj χ^2/df fit statistic at 1.0

	Adj χ^2/df		
Sample Size	Singles	Doubles	Triples
	10 Items		
500	0	3	4
1000	0	3	5
2000	0	6	7
	20 Items		
500	0	4	6
1000	0	4	5
2000	0	7	7
	40 Items		
500	0	6	6
1000	0	5	8
2000	0	12	14

Note. Percentages out of 100 samples and across all items. Adj χ^2/df were computed without cross validating.

Table 6

Overall Type I Error percentages for the Adj χ^2/df fit statistic at 2.0

	Adj χ^2/df		
Sample Size	Singles	Doubles	Triples
	10 Items		
500	0	2	2
1000	0	1	2
2000	0	1	1
	20 Items		
500	0	3	3
1000	0	1	2
2000	0	1	0
	40 Items		
500	0	4	4
1000	0	2	2
2000	0	3	2

Note. Percentages out of 100 samples and across all items. Adj χ^2/df were computed without cross validating.

Table 7

Overall Type I Error percentages for the Adj χ^2/df fit statistic at 4.0

	Adj χ^2/df		
Sample Size	Singles	Doubles	Triples
	10 Items		
500	0	1	1
1000	0	0	0
2000	0	0	0
	20 Items		
500	0	1	1
1000	0	0	0
2000	0	0	0
	40 Items		
500	0	1	1
1000	0	0	0
2000	0	0	0

Note. Percentages out of 100 samples and across all items. Adj χ^2/df were computed without cross validating.

Table 8

Overall power estimates for the Adj χ^2/df fit statistic at 1.0

Number of items	Sample Size	Adj χ^2/df		
		Singles	Doubles	Triples
20% items w/ LD				
10 items	N = 500	8	33	52
	N = 1000	9	58	82
	N = 2000	6	72	90
20 items	N = 500	0	25	40
	N = 1000	5	60	80
	N = 2000	3	69	87
40 items	N = 500	0	30	42
	N = 1000	5	64	86
	N = 2000	2	67	85
40% items w/ LD				
10 items	N = 500	19	53	78
	N = 1000	19	54	80
	N = 2000	17	80	97
20 items	N = 500	8	42	73
	N = 1000	16	56	83
	N = 2000	20	79	97
40 items	N = 500	6	60	83
	N = 1000	12	72	91
	N = 2000	18	89	99
Items w/ restricted θ range				
10 items	N = 500	0	3	5
	N = 1000	0	4	6
	N = 2000	0	4	6
20 items	N = 500	1	5	6
	N = 1000	0	5	6
	N = 2000	0	5	9
40 items	N = 500	0	5	6
	N = 1000	0	5	9
	N = 2000	0	5	9
GGUM Generated				
10 items	N = 500	3	19	30

	N = 1000	1	16	27
	N = 2000	0	19	33
20 items	N = 500	2	23	36
	N = 1000	2	26	42
	N = 2000	4	34	49
40 items	N = 500	4	15	21
	N = 1000	2	18	30
	N = 2000	1	21	36

Note. Percentages out of 100 samples and across all misfitting items. GGUM is the generalized graded unfolding model. Adj χ^2/df were computed without cross validating.

Table 9

Overall power estimates for the Adj χ^2/df fit statistic at 2.0

Number of items	Sample Size	Adj χ^2/df		
		Singles	Doubles	Triples
20% items w/ LD				
10 items	N = 500	7	29	45
	N = 1000	7	51	73
	N = 2000	4	59	79
20 items	N = 500	0	20	32
	N = 1000	3	55	77
	N = 2000	2	52	72
40 items	N = 500	0	26	35
	N = 1000	4	56	77
	N = 2000	1	49	60
40% items w/ LD				
10 items	N = 500	18	49	73
	N = 1000	16	48	72
	N = 2000	11	36	94
20 items	N = 500	6	37	68
	N = 1000	14	48	75
	N = 2000	17	69	94
40 items	N = 500	4	56	78
	N = 1000	13	66	87
	N = 2000	13	82	97
Items w/ restricted θ range				
10 items	N = 500	0	2	2
	N = 1000	0	1	2
	N = 2000	0	1	1
20 items	N = 500	1	4	4
	N = 1000	0	2	2
	N = 2000	0	1	1
40 items	N = 500	0	3	3
	N = 1000	0	3	3
	N = 2000	0	1	1
GGUM Generated				
10 items	N = 500	3	16	24

	N = 1000	1	11	15
	N = 2000	0	9	11
20 items	N = 500	2	19	29
	N = 1000	2	20	30
	N = 2000	3	22	31
40 items	N = 500	3	18	25
	N = 1000	2	13	20
	N = 2000	1	12	16

Note. Percentages out of 100 samples and across all misfitting items. GGUM is the generalized graded unfolding model. Adj χ^2/df were computed without cross validating.

Table 10
Overall power estimates for the Adj χ^2/df fit statistic at 4.0

Number of items	Sample Size	Adj χ^2/df		
		Singles	Doubles	Triples
20% items w/ LD				
10 items	N = 500	6	23	35
	N = 1000	6	38	58
	N = 2000	2	40	58
20 items	N = 500	0	13	22
	N = 1000	2	40	60
	N = 2000	1	33	46
40 items	N = 500	0	21	26
	N = 1000	2	42	59
	N = 2000	0	28	32
40% items w/ LD				
10 items	N = 500	15	42	65
	N = 1000	12	38	58
	N = 2000	6	58	84
20 items	N = 500	5	30	58
	N = 1000	10	35	61
	N = 2000	11	53	84
40 items	N = 500	3	48	69
	N = 1000	9	55	75
	N = 2000	7	72	91
Items w/ restricted θ range				
10 items	N = 500	0	3	5
	N = 1000	0	0	0
	N = 2000	0	0	0
20 items	N = 500	1	2	2
	N = 1000	0	0	0
	N = 2000	0	0	0
40 items	N = 500	0	1	1
	N = 1000	0	1	0
	N = 2000	0	0	0
GGUM Generated				
10 items	N = 500	3	12	16

	N = 1000	1	6	6
	N = 2000	0	2	1
20 items	N = 500	2	14	19
	N = 1000	2	12	15
	N = 2000	2	10	12
40 items	N = 500	2	12	20
	N = 1000	2	8	11
	N = 2000	1	5	5

Note. Percentages out of 100 samples and across all misfitting items. GGUM is the generalized graded unfolding model. Adj χ^2/df were computed without cross validating.

Figures

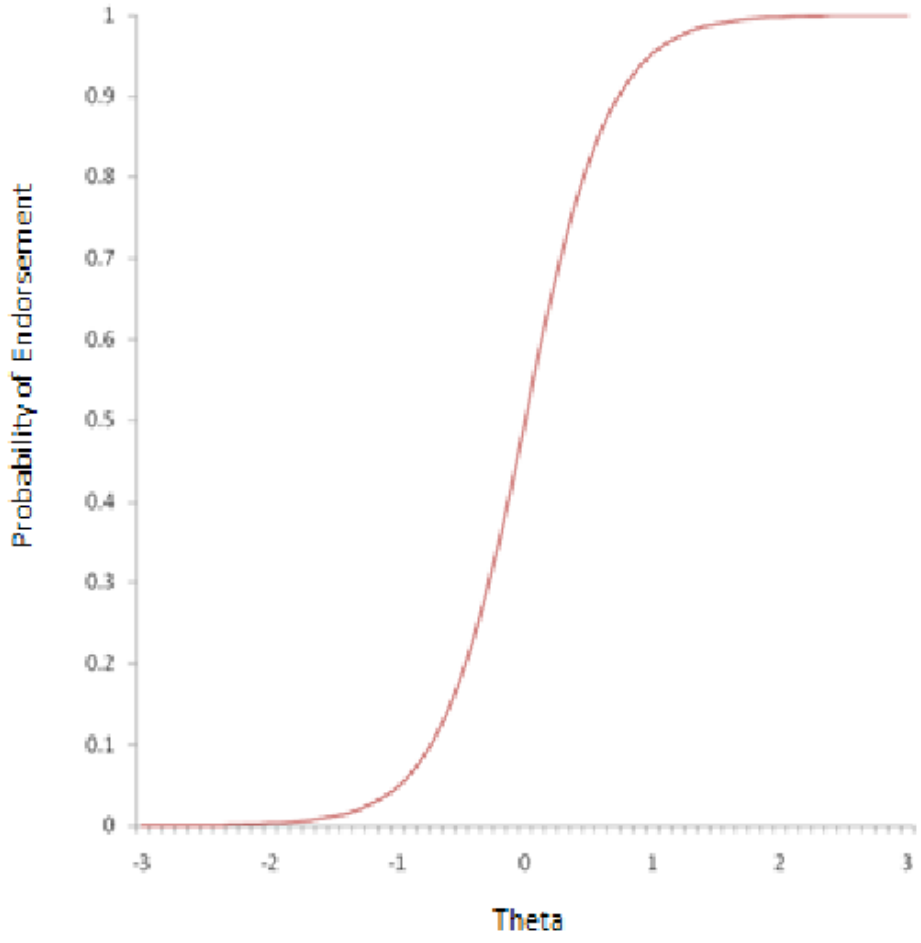

Figure 1. Item characteristic curve for an item with a difficulty of 0, and a discrimination of 5.0.

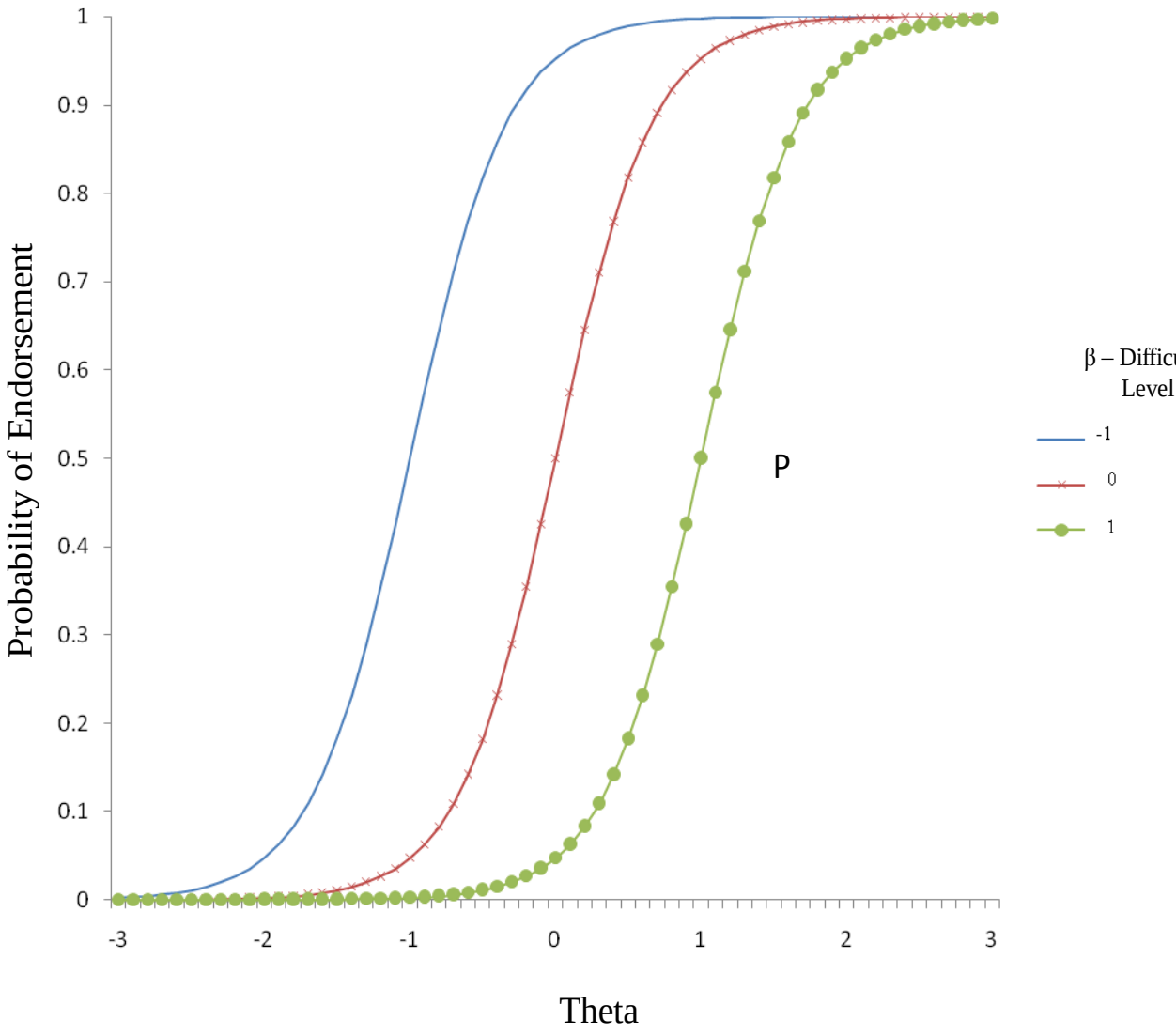

Figure 2. Three items with the same discrimination values, but with three different difficulty levels of -1.0, 0, and 1.0.

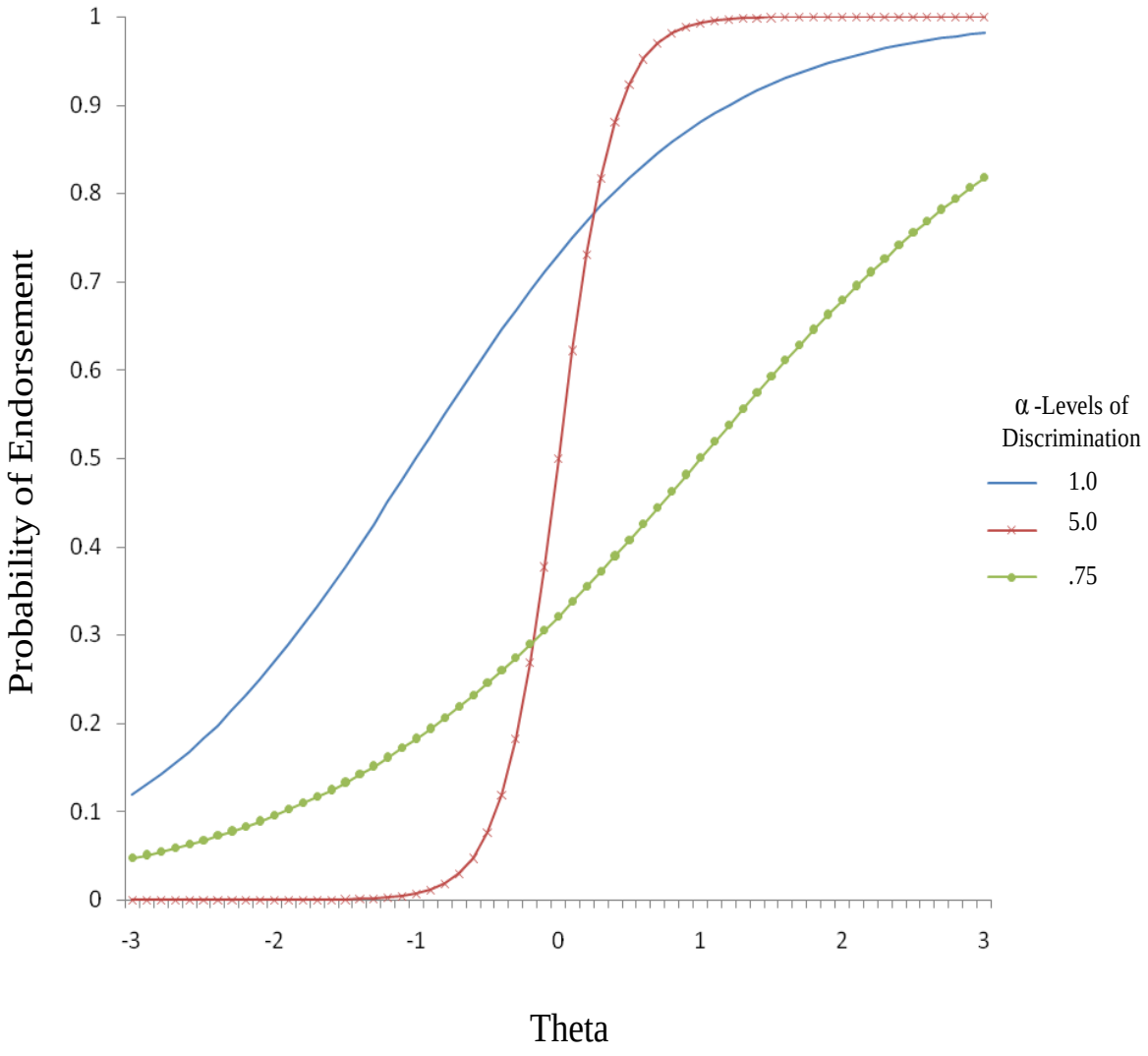

Figure 3. Allowing the discrimination parameter to be freely estimated. By doing this the slopes of the individual IRFs can differ substantially.

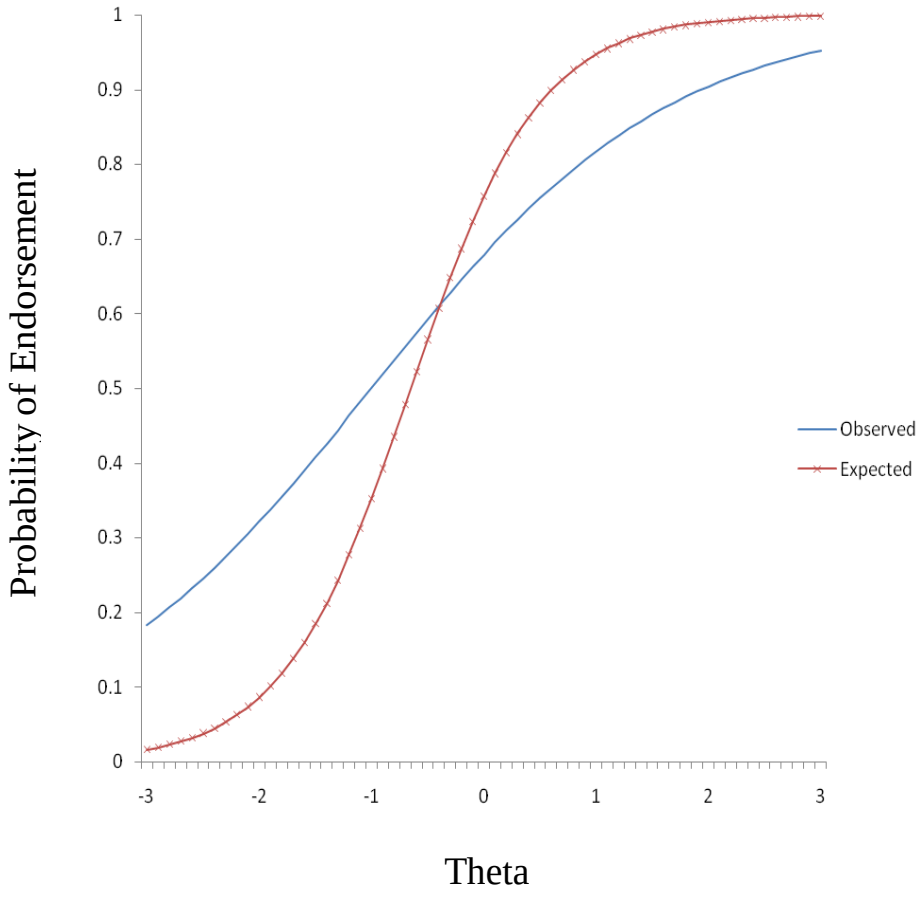

Figure 4. Example of when an empirical item response function (IRF) is above the estimated IRF at low trait levels and is below it at high trait levels.

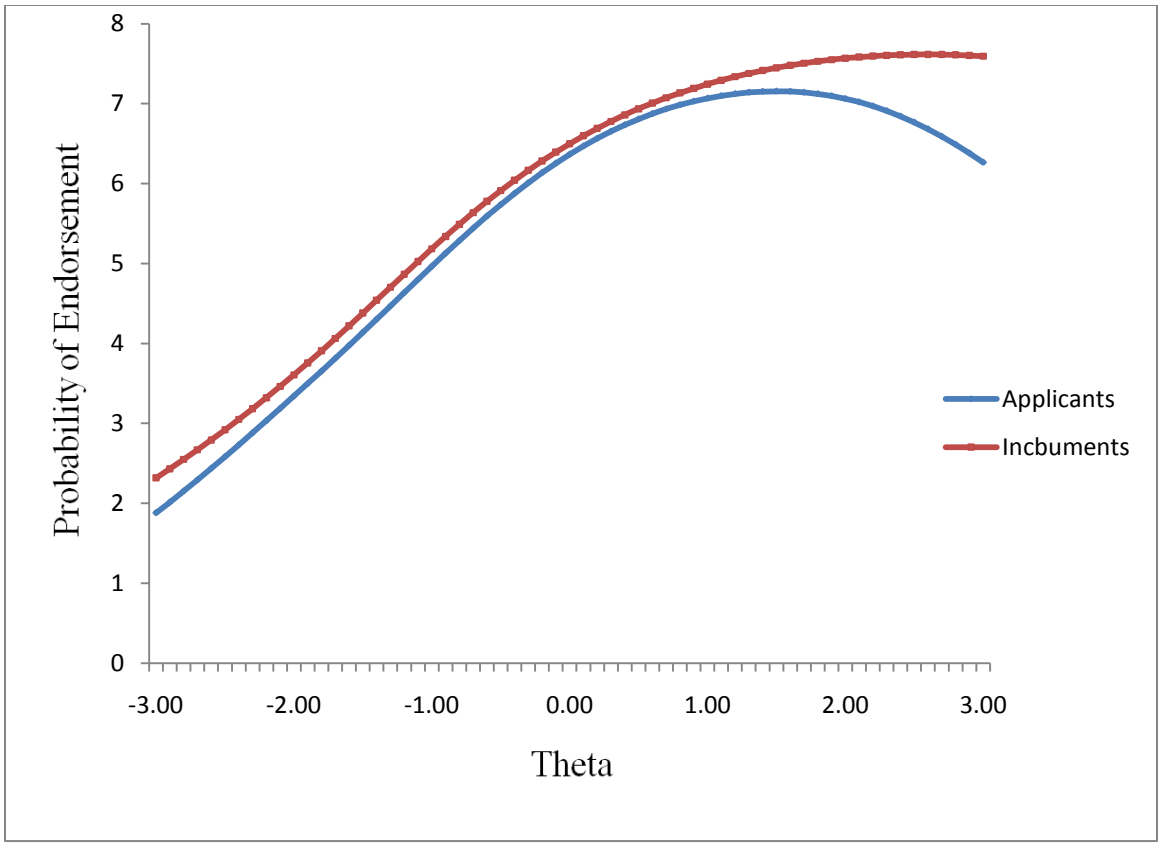

Figure 5. Different responding patterns between applicants and incumbents on a personality assessment. This demonstrates an ideal point responding process.